RISK SIMULATOR

User Manual

Johnathan Mun, Ph.D., MBA, MS, BS, CRM, FRM, CFC, MIFC
Real Options Valuation, Inc.

REAL OPTIONS VALUATION, INC.

This manual, and the software described in it, are furnished under license and may only be used or copied in accordance with the terms of the end user license agreement. Information in this document is provided for informational purposes only, is subject to change without notice, and does not represent a commitment as to merchantability or fitness for a particular purpose by Real Options Valuation, Inc. No part of this manual may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording, for any purpose without the express written permission of Real Options Valuation, Inc. Materials based on copyrighted publications by Dr. Johnathan Mun, Ph.D., MBA, MS, BS, CRM, CFC, FRM, MIFC, Founder and CEO, Real Options Valuation, Inc., and creator of the software. Written, designed, and published in the United States of America. Microsoft® is a registered trademark of Microsoft Corporation in the U.S. and other countries. Other product names mentioned herein may be trademarks and/or registered trademarks of the respective holders.

© Copyright 2005-2012 Dr. Johnathan Mun. All rights reserved.

Real Options Valuation, Inc.

4101F Dublin Blvd., Ste. 425

Dublin, California 94568 U.S.A.

Phone 925.271.4438 • Fax 925.369.0450

admin@realoptionsvaluation.com

www.risksimulator.com

www.realoptionsvaluation.com

Table of Contents

1. INTRODUCTION TO RISK SIMULATOR 1

1.1 Welcome to Risk Simulator..... 1

1.2 Installation Requirements and Procedures..... 2

1.3 Licensing..... 2

1.4 What's New in Version 2012..... 5

1.4.1 General Capabilities 5

1.4.2 Simulation Module 6

1.4.3 Forecasting Module..... 7

1.4.4 Optimization Module 7

1.4.5 Analytical Tools Module 8

1.4.6 Statistics and BizStats Module 9

2. MONTE CARLO RISK SIMULATION..... 11

2.1 What Is Monte Carlo Simulation? 11

2.2 Getting Started with Risk Simulator..... 12

2.2.1 A High-Level Overview of the Software..... 12

2.2.2 Running a Monte Carlo Simulation 13

2.2.3 Forecast Chart Tabs..... 21

2.3.4 Using Forecast Charts and Confidence Intervals 23

2.3 Correlations and Precision Control 26

2.3.1 The Basics of Correlations 26

2.3.2 Applying Correlations in Risk Simulator..... 27

2.3.3 The Effects of Correlations in Monte Carlo Simulation..... 27

2.3.4 Precision and Error Control 29

2.3.5 Understanding the Forecast Statistics 31

2.3.6 Understanding Probability Distributions for Monte Carlo Simulation 35

2.4 Discrete Distributions 38

2.5 Continuous Distributions 45

3. FORECASTING 64

3.1 Different Types of Forecasting Techniques 65

3.2 Running the Forecasting Tool in Risk Simulator 68

3.3 Time-Series Analysis 69

3.4 Multivariate Regression 72

3.5 Stochastic Forecasting 76

3.6 Nonlinear Extrapolation 78

3.7 Box-Jenkins ARIMA Advanced Time-Series 80

3.8 AUTO-ARIMA (Box-Jenkins ARIMA Advanced Time-Series) 85

3.9 Basic Econometrics 86

3.10 J-S Curve Forecasts 89

3.11 GARCH Volatility Forecasts 90

3.12 Markov Chains 93

3.13 Limited Dependent Variables: Logit, Probit, Tobit (Maximum Likelihood Estimation) 94

3.14 Spline (Cubic Spline Interpolation and Extrapolation) 97

4. OPTIMIZATION 99

4.1 Optimization Methodologies 99

4.2 Optimization with Continuous Decision Variables 101

4.3 Optimization with Discrete Integer Variables 106

4.4 Efficient Frontier and Advanced Optimization Settings *109*

4.5 Stochastic Optimization *111*

5. RISK SIMULATOR ANALYTICAL TOOLS 116

5.1 Tornado and Sensitivity Tools in Simulation *116*

5.2 Sensitivity Analysis *123*

5.3 Distributional Fitting: Single Variable and Multiple Variables *127*

5.4 Bootstrap Simulation *131*

5.5 Hypothesis Testing *133*

5.6 Data Extraction and Saving Simulation Results *135*

5.7 Create Report *136*

5.8 Regression and Forecasting Diagnostic Tool *137*

5.9 Statistical Analysis Tool *144*

5.10 Distributional Analysis Tool *147*

5.11 Scenario Analysis Tool *150*

5.12 Segmentation Clustering Tool *153*

5.13 RISK SIMULATOR 2011/2012 NEW TOOLS *154*

5.14 Random Number Generation, Monte Carlo vs. Latin Hypercube, and Correlation Copulas .. *154*

5.15 Deseasonalizing and Detrending Data *155*

5.16 Principal Component Analysis *157*

5.17 Structural Break Analysis *157*

5.18 Trendline Forecasts *159*

5.19 Model Checking Tool *159*

5.20 Percentile Distributional Fitting Tool *160*

5.21 Distribution Charts and Tables: Probability Distribution Tool *162*

5.22 ROV BizStats *165*

5.23 Neural Network and Combinatorial Fuzzy Logic Forecasting Methodologies *169*

5.24 Optimizer Goal Seek *171*

5.25 Single Variable Optimizer *172*

5.26 Genetic Algorithm Optimization *173*

5.27 ROV Decision Tree Module *174*

5.27.1 Decision Tree 174

5.27.2 Simulation Modeling 176

5.27.3 Bayes Analysis 177

5.27.4 Expected Value of Perfect Information, MINIMAX and MAXIMIN Analysis, Risk Profiles, and Value of Imperfect Information 177

5.27.5 Sensitivity 178

5.27.6 Scenario Tables 178

5.27.7 Utility Function Generation 178

6. Helpful Tips and Techniques 183

TIPS: Assumptions (Set Input Assumption User Interface) 183

TIPS: Copy and Paste 183

TIPS: Correlations 184

TIPS: Data Diagnostics and Statistical Analysis 184

TIPS: Distributional Analysis, Charts and Probability Tables 184

TIPS: Efficient Frontier 185

TIPS: Forecast Cells 185

TIPS: Forecast Charts 185

TIPS: Forecasting 185

TIPS: Forecasting: ARIMA....185

TIPS: Forecasting: Basic Econometrics....186

TIPS: Forecasting: Logit, Probit, and Tobit....186

TIPS: Forecasting: Stochastic Processes....186

TIPS: Forecasting: Trendlines....186

TIPS: Function Calls....186

TIPS: Getting Started Exercises and Getting Started Videos....186

TIPS: Hardware ID....187

TIPS: Latin Hypercube Sampling (LHS) vs. Monte Carlo Simulation (MCS)....187

TIPS: Online Resources....187

TIPS: Optimization....187

TIPS: Profiles....188

TIPS: Right-Click Shortcut and Other Shortcut Keys....188

TIPS: Save....188

TIPS: Sampling and Simulation Techniques....189

TIPS: Software Development Kit (SDK) and DLL Libraries....189

TIPS: Starting Risk Simulator with Excel....189

TIPS: Super Speed Simulation....189

TIPS: Tornado Analysis....189

TIPS: Troubleshooter....190

INDEX....191

1

1. INTRODUCTION TO RISK SIMULATOR

1.1 Welcome to Risk Simulator

The Risk Simulator is a Monte Carlo simulation, Forecasting, and Optimization software. The software is written in Microsoft .NET C# and functions together with Excel as an add-in. This software is also compatible and often used with the Real Options Super Lattice Solver (SLS) software and Employee Stock Options Valuation Toolkit (ESOV) software, also developed by Real Options Valuation, Inc. Note that although we attempt to be thorough in this user manual, the manual is absolutely not a substitute for the Training DVD, live training courses, and books written by the software's creator (e.g., Dr. Johnathan Mun's Real Options Analysis, 2nd Edition, Wiley Finance, 2005; Modeling Risk: Applying Monte Carlo Simulation, Real Options Analysis, Forecasting, and Optimization, 2nd Edition, Wiley Finance, 2010; and Valuing Employee Stock Options (2004 FAS 123R), Wiley Finance, 2004). Please visit our website at www.realoptionsvaluation.com for more information about these items.

The Risk Simulator software has the following modules:

- Monte Carlo Simulation (runs parametric and nonparametric simulation of 42 probability distributions with different simulation profiles, truncated and correlated simulations, customizable distributions, precision and error-controlled simulations, and many other algorithms)
- Forecasting (runs Box-Jenkins ARIMA, multiple regression, nonlinear extrapolation, stochastic processes, and time-series analysis)
- Optimization Under Uncertainty (runs optimizations using discrete integer and continuous variables for portfolio and project optimization with and without simulation)
- Modeling and Analytical Tools (runs tornado, spider, and sensitivity analysis, as well as bootstrap simulation, hypothesis testing, distributional fitting, etc.)
- ROV BizStats (over 130 business statistics and analytical models)
- ROV Decision Tree (decision tree models, Monte Carlo risk simulation on decision trees, sensitivity analysis, scenario analysis, Bayesian joint and posterior probability updating, expected value of information, MINIMAX, MAXIMIN, risk profiles)

Real Options SLS software is used for computing simple and complex options and includes the ability to create customizable option models. This software has the following modules:

- Single Asset SLS (for solving abandonment, chooser, contraction, deferment, and expansion options, as well as for solving customized options)
- Multiple Asset and Multiple Phase SLS (for solving multiphase sequential options, options with multiple underlying assets and phases, combination of multiphase sequential with abandonment, chooser, contraction, deferment, expansion, and switching options; it can also be used to solve customized options)
- Multinomial SLS (for solving trinomial mean-reverting options, quadranomial jump-diffusion options, and pentanomial rainbow options)
- Excel Add-In Functions (for solving all the above options plus closed-form models and customized options in an Excel-based environment)

1.2 Installation Requirements and Procedures

To install the software, follow the on-screen instructions. The minimum requirements for this software are:

- Pentium IV processor or later (dual core recommended)
- Windows XP, Vista, Windows 7, Windows 8, or later
- Microsoft Excel XP, 2003, 2007, 2010, or later
- Microsoft .NET Framework 2.0 or later (versions 3.0, 3.5, and so forth)
- 500 MB free space
- 2GB RAM minimum (2–4GB recommended)
- Administrative rights to install software

Most new computers come with Microsoft .NET Framework 2.0/3.0 already installed. However, if an error message pertaining to requiring .NET Framework occurs during the installation of Risk Simulator, exit the installation. Then, install the relevant .NET Framework software included in the CD (choose your own language). Complete the .NET installation, restart the computer, and then reinstall the Risk Simulator software.

There is a default 10-day trial license file that comes with the software. To obtain a full corporate license, please contact Real Options Valuation, Inc., at *admin@realoptionsvaluation.com* or call +1 (925) 271-4438, or visit our website at *www.realoptionsvaluation.com*. Please visit this website and click on DOWNLOAD to obtain the latest software release, or click on the FAQ link to obtain any updated information on licensing or installation issues and fixes.

1.3 Licensing

If you have installed the software and have purchased a full license to use the software, you will need to e-mail us your Hardware ID so that we can generate a license file for you. Follow the instructions below:

- Start Excel XP/2003/2007/2010, click on the License icon or *Risk Simulator | License* and copy down and e-mail your 11 to 20 digit and alphanumeric HARDWARE ID that starts with the prefix "RS" (you can also select the Hardware ID and do a right-click copy or click on the e-mail Hardware ID link) to *admin@realoptionsvaluation.com*. Once we have obtained this ID, a newly generated permanent license will be e-mailed to you. Once you obtain this license file, simply save it to your hard drive (if it is a zipped file, first unzip its contents and save them to your hard drive). Start Excel, click *on Risk Simulator | License* or click on the License icon and click on Install License and point to this new license file. Restart Excel and you are done. The entire process will take less than a minute and you will be fully licensed.

- Once installation is complete, start Microsoft Excel and if the installation was successful, you should see an additional "Risk Simulator" item on the menu bar in Excel XP/2003 or under the new icon group in Excel 2007/2010, and a new icon bar on Excel as seen in Figure 1.1. In addition, a splash screen will appear as seen in Figure 1.2, indicating that the software is functioning and loaded into Excel. Figure 1.3 also shows the Risk Simulator toolbar. If these items exist in Excel, you are now ready to start using the software. The remainder of this user manual provides step-by-step instructions for using the software.

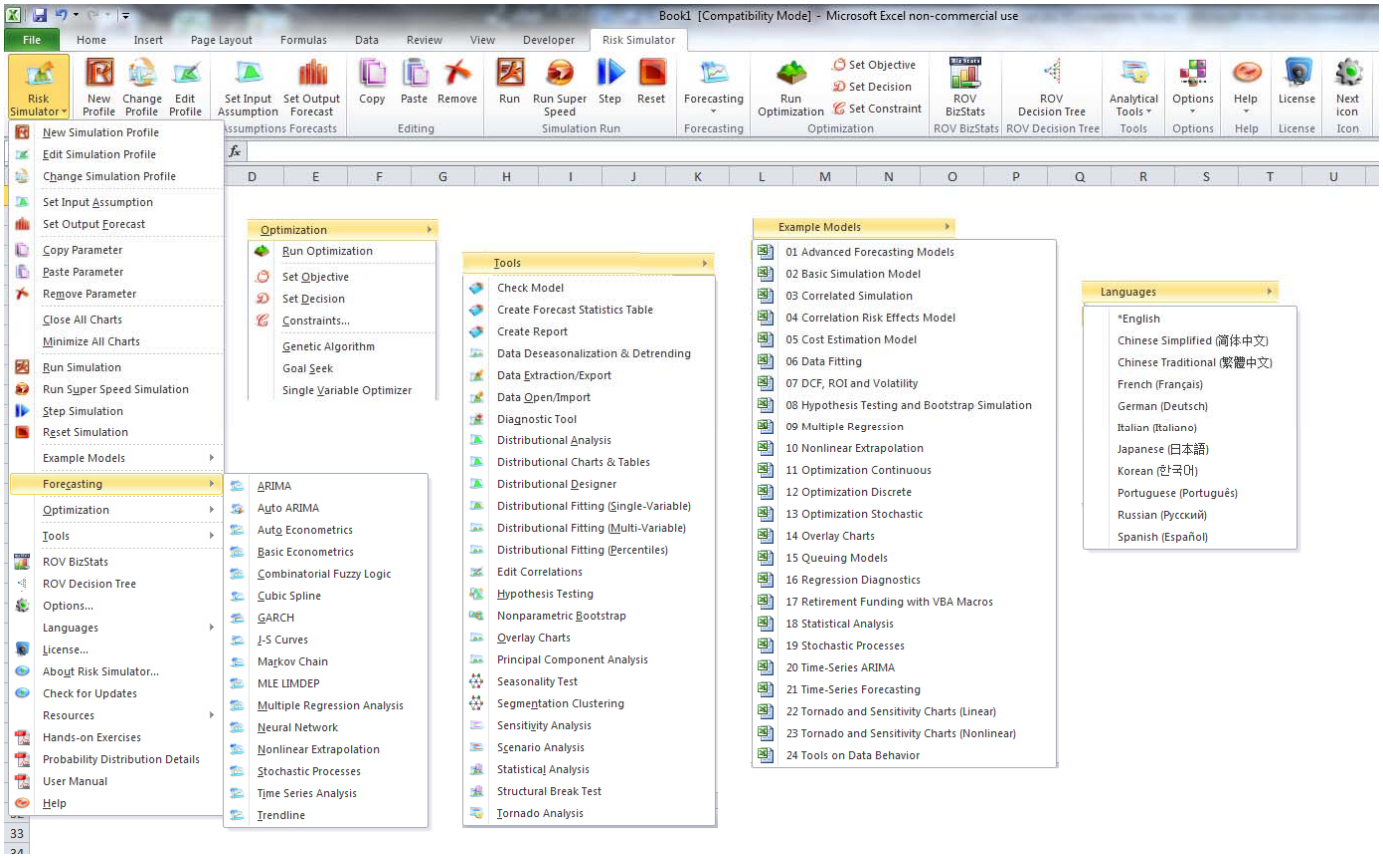


Figure 1.1 – Risk Simulator Menu and Icon Bar in Excel 2007/2010

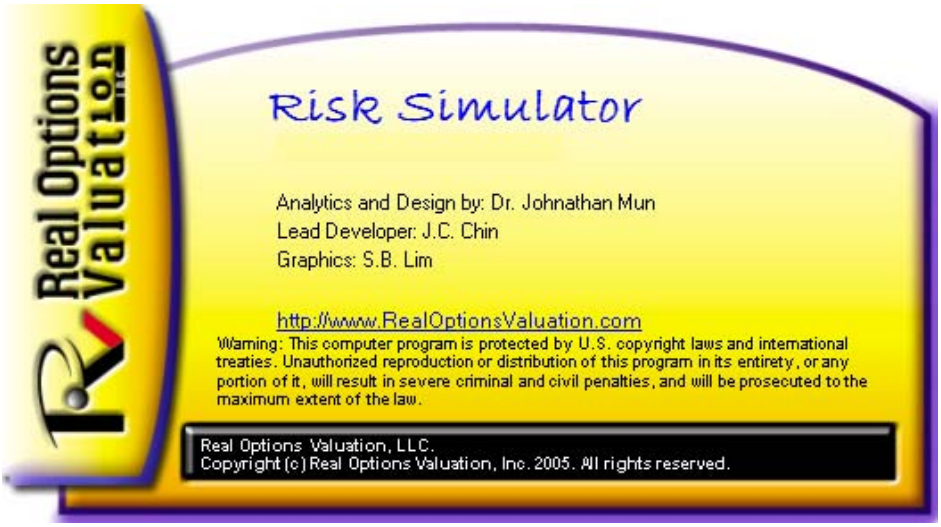


Figure 1.2 – Risk Simulator Splash Screen

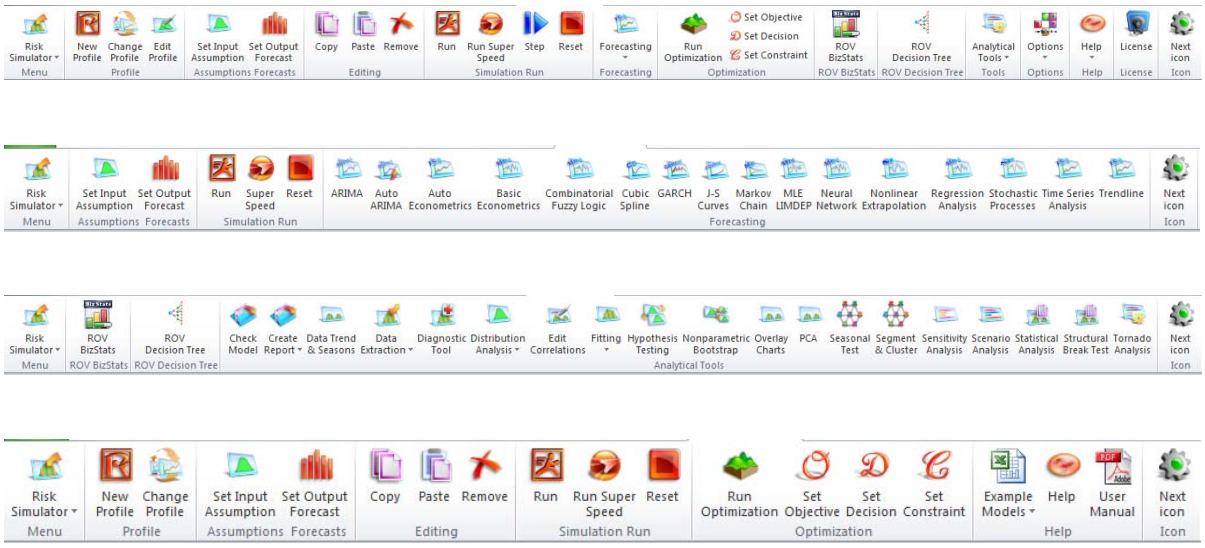


Figure 1.3 – Risk Simulator Icon Toolbars in Excel 2007/2010

1.4 What's New in Version 2012

The following lists the main capabilities of Risk Simulator, where the highlighted items indicate the latest additions to version 2011/2012.

1.4.1 General Capabilities

1. Available in 11 languages—English, French, German, Italian, Japanese, Korean, Portuguese, Russian, Spanish, Simplified Chinese, and Traditional Chinese.
2. ROV Decision Tree module is included in the latest version and is used to create and value decision tree models. Additional advanced methodologies and analytics are also included:
    - Decision Tree Models
    - Monte Carlo risk simulation
    - Sensitivity Analysis
    - Scenario Analysis
    - Bayesian (Joint and Posterior Probability Updating)
    - Expected Value of Information
    - MINIMAX
    - MAXIMIN
    - Risk Profiles
3. Books—analytical theory, application, and case studies are supported by 10 books.
4. Commented Cells—turn cell comments on or off and decide if you wish to show cell comments on all input assumptions, output forecasts, and decision variables.
5. Detailed Example Models—24 example models in Risk Simulator and over 300 models in Modeling Toolkit.
6. Detailed Reports—all analyses come with detailed reports.
7. Detailed User Manual—step-by-step user manual.
8. Flexible Licensing—certain functionalities can be turned on or off to allow you to customize your risk analysis experience. For instance, if you are only interested in the forecasting tools in Risk Simulator, you may be able to obtain a special license that activates only the forecasting tools and leaves the other modules deactivated, thereby saving some costs on the software.
9. Flexible Requirements—works in Window 7, Vista, and XP; integrates with Excel 2010, 2007, 2003; and works in MAC operating systems running virtual machines.
10. Fully customizable colors and charts—tilt, 3D, color, chart type, and much more!
11. Hands-on Exercises—detailed step-by-step guide to running Risk Simulator, including guides on interpreting the results.
12. Multiple Cell Copy and Paste—allows assumptions, decision variables, and forecasts to be copied and pasted.

13. Profiling—allows multiple profiles to be created in a single model (different scenarios of simulation models can be created, duplicated, edited, and run in a single model).
14. Revised Icons in Excel 2007/2010—a completely reworked icon toolbar that is more intuitive and user friendly. There are four sets of icons that fit most screen resolutions (1280 x 760 and above).
15. Right-Click Shortcuts—access all of Risk Simulator's tools and menus using a mouse right-click.
16. ROV Software Integration—works well with other ROV software including Real Options SLS, Modeling Toolkit, Basel Toolkit, ROV Compiler, ROV Extractor and Evaluator, ROV Modeler, ROV Valuator, ROV Optimizer, ROV Dashboard, ESO Valuation Toolkit, and others!
17. RS Functions in Excel—insert RS functions for setting assumptions and forecasts, and right-click support in Excel.
18. Troubleshooter—allows you to re-enable the software, check for your system requirements, obtain the Hardware ID, and others.
19. Turbo Speed Analysis—runs forecasts and other analyses tools at blazingly fast speeds (enhanced in version 5.2). The analyses and results remain the same but are now computed very quickly; reports are generated very quickly as well.
20. Web Resources, Case Studies, and Videos—download free models, getting-started videos, case studies, whitepapers, and other materials from our website.

1.4.2 Simulation Module

21. 6 random number generators—ROV Advanced Subtractive Generator, Subtractive Random Shuffle Generator, Long Period Shuffle Generator, Portable Random Shuffle Generator, Quick IEEE Hex Generator, and Basic Minimal Portable Generator.
22. 2 sampling methods—Monte Carlo and Latin Hypercube.
23. 3 Correlation Copulas—applying Normal Copula, T Copula, and Quasi-Normal Copula for correlated simulations.
24. 42 probability distributions—arcsine, Bernoulli, beta, beta 3, beta 4, binomial, Cauchy, chi-square, cosine, custom, discrete uniform, double log, Erlang, exponential, exponential 2, F distribution, gamma, geometric, Gumbel max, Gumbel min, hypergeometric, Laplace, logistic, lognormal (arithmetic) and lognormal (log), lognormal 3 (arithmetic) and lognormal 3 (log), negative binomial, normal, parabolic, Pareto, Pascal, Pearson V, Pearson VI, PERT, Poisson, power, power 3, Rayleigh, t and t2, triangular, uniform, Weibull, Weibull 3.
25. Alternate Parameters—using percentiles as an alternate way of inputting parameters.
26. Custom Nonparametric Distribution—make your own distributions for running historical simulations, and applying the Delphi method.
27. Distribution Truncation—enabling data boundaries.
28. Excel Functions—set assumptions and forecasts using functions inside Excel
29. Multidimensional Simulation—simulation of uncertain input parameters.

30. Precision Control—determines if the number of simulation trials run is sufficient.
31. Super Speed Simulation—runs 100,000 trials in a few seconds.

1.4.3 Forecasting Module

32. ARIMA—autoregressive integrated moving average models ARIMA (P,D,Q).
33. Auto ARIMA—runs the most common combinations of ARIMA to find the best-fitting model.
34. Auto Econometrics—runs thousands of model combinations and permutations to obtain the best-fitting model for existing data (linear, nonlinear, interacting, lag, leads, rate, difference).
35. Basic Econometrics—econometric and linear/nonlinear and interacting regression models.
36. Combinatorial Fuzzy Logic Forecasts—time-series forecast methods
37. Cubic Spline—nonlinear interpolation and extrapolation.
38. GARCH—volatility projections using generalized autoregressive conditional heteroskedasticity models: GARCH, GARCH-M, TGARCH, TGARCH-M, EGARCH, EGARCH-T, GJR-GARCH, and GJR-TGARCH.
39. J-Curve—exponential J curves.
40. Limited Dependent Variables—Logit, Probit, and Tobit.
41. Markov Chains—two competing elements over time and market share predictions.
42. Multiple Regression—regular linear and nonlinear regression, with stepwise methodologies (forward, backward, correlation, forward-backward).
43. Neural Network Forecasts—linear, nonlinear logistic, hyperbolic tangent, and cosine
44. Nonlinear Extrapolation—nonlinear time-series forecasting.
45. S Curve—logistic S curves.
46. Time-Series Analysis—8 time-series decomposition models for predicting levels, trends, and seasonalities.
47. Trendlines—forecasting and fitting using linear, nonlinear polynomial, power, logarithmic, exponential, and moving averages with goodness of fit.

1.4.4 Optimization Module

48. Linear Optimization—multiphasic optimization and general linear optimization.
49. Nonlinear Optimization—detailed results including Hessian matrices, LaGrange functions, and more.
50. Static Optimization—quick runs for continuous, integers, and binary optimizations.
51. Dynamic Optimization—simulation with optimization.
52. Stochastic Optimization—quadratic, tangential, central, forward, and convergence criteria.

53. Efficient Frontier—combinations of stochastic and dynamic optimizations on multivariate efficient frontiers.
54. Genetic Algorithms—used for a variety of optimization problems.
55. Multiphasic Optimization—testing for local versus global optimum allowing better control over how the optimization is run, and increases the accuracy and dependency of the results.
56. Percentiles and Conditional Means—additional statistics for stochastic optimization, including percentiles as well as conditional means, which are critical in computing conditional value at risk measures.
57. Search Algorithm—simple, fast, and efficient search algorithms for basic single decision variable and goal seek applications.
58. Super Speed Simulation in Dynamic and Stochastic Optimization—runs simulation at super speed while integrated with optimization.

1.4.5 Analytical Tools Module

59. Check Model—tests for the most common mistakes in your model.
60. Correlation Editor—allows large correlation matrices to be directly entered and edited.
61. Create Report—automates report generation of assumptions and forecasts in a model.
62. Create Statistics Report—generates comparative report of all forecast statistics.
63. Data Diagnostics—runs tests on heteroskedasticity, micronumerosity, outliers, nonlinearity, autocorrelation, normality, sphericity, nonstationarity, multicollinearity, and correlations.
64. Data Extraction and Export—extracts data to Excel or flat text files and Risk Sim files, runs statistical reports and forecast result reports.
65. Data Open and Import—retrieves previous simulation run results.
66. Deseasonalization and Detrending—deseasonalizes and detrends your data.
67. Distributional Analysis—computes exact PDF, CDF, and ICDF of all 42 distributions and generates probability tables.
68. Distributional Designer—allows you to create custom distributions.
69. Distributional Fitting (Multiple)— runs multiple variables simultaneously, accounts for correlations and correlation significance.
70. Distributional Fitting (Single)—Kolmogorov-Smirnov and chi-square tests on continuous distributions, complete with reports and distributional assumptions.
71. Hypothesis Testing—tests if two forecasts are statistically similar or different.
72. Nonparametric Bootstrap—simulation of the statistics to obtain the precision and accuracy of the results.
73. Overlay Charts—fully customizable overlay charts of assumptions and forecasts together (CDF, PDF, 2D/3D chart types).
74. Principal Component Analysis—tests the best predictor variables and ways to reduce the data array.

75. Scenario Analysis—hundreds and thousands of static two-dimensional scenarios.
76. Seasonality Test—tests for various seasonality lags.
77. Segmentation Clustering—groups data into statistical clusters for segmenting your data.
78. Sensitivity Analysis—dynamic sensitivity (simultaneous analysis).
79. Structural Break Test—tests if your time-series data has statistical structural breaks.
80. Tornado Analysis—static perturbation of sensitivities, spider and tornado analysis, and scenario tables.

1.4.6 Statistics and BizStats Module

81. Percentile Distributional Fitting—using percentiles and optimization to find the best-fitting distribution.
82. Probability Distributions' Charts and Tables—run 45 probability distributions, their four moments, CDF, ICDF, PDF, charts, and overlay multiple distributional charts, and generate probability distribution tables.
83. Statistical Analysis—descriptive statistics, distributional fitting, histograms, charts, nonlinear extrapolation, normality test, stochastic parameters estimation, time-series forecasting, trendline projections, etc.
84. ROV BIZSTATS—over 130 business statistics and analytical models:

    Absolute Values, ANOVA: Randomized Blocks Multiple Treatments, ANOVA: Single Factor Multiple Treatments, ANOVA: Two Way Analysis, ARIMA, Auto ARIMA, Autocorrelation and Partial Autocorrelation, Autoeconometrics (Detailed), Autoeconometrics (Quick), Average, Combinatorial Fuzzy Logic Forecasting, Control Chart: C, Control Chart: NP, Control Chart: P, Control Chart: R, Control Chart: U, Control Chart: X, Control Chart: XMR, Correlation, Correlation (Linear, Nonlinear), Count, Covariance, Cubic Spline, Custom Econometric Model, Data Descriptive Statistics, Deseasonalize, Difference, Distributional Fitting, Exponential J Curve, GARCH, Heteroskedasticity, Lag, Lead, Limited Dependent Variables (Logit), Limited Dependent Variables (Probit), Limited Dependent Variables (Tobit), Linear Interpolation, Linear Regression, LN, Log, Logistic S Curve, Markov Chain, Max, Median, Min, Mode, Neural Network, Nonlinear Regression, Nonparametric: Chi-Square Goodness of Fit, Nonparametric: Chi-Square Independence, Nonparametric: Chi-Square Population Variance, Nonparametric: Friedman's Test, Nonparametric: Kruskal-Wallis Test, Nonparametric: Lilliefors Test, Nonparametric: Runs Test, Nonparametric: Wilcoxon Signed-Rank (One Var), Nonparametric: Wilcoxon Signed-Rank (Two Var), Parametric: One Variable (T) Mean, Parametric: One Variable (Z) Mean, Parametric: One Variable (Z) Proportion, Parametric: Two Variable (F) Variances, Parametric: Two Variable (T) Dependent Means, Parametric: Two Variable (T) Independent Equal Variance, Parametric: Two Variable (T) Independent Unequal Variance, Parametric: Two Variable (Z) Independent Means, Parametric: Two Variable (Z) Independent Proportions, Power, Principal Component Analysis, Rank Ascending, Rank Descending, Relative LN Returns, Relative Returns, Seasonality, Segmentation Clustering, Semi-Standard Deviation (Lower), Semi-Standard Deviation (Upper), Standard 2D Area, Standard 2D Bar, Standard 2D Line, Standard 2D Point, Standard 2D Scatter, Standard 3D Area, Standard 3D Bar, Standard

3D Line, Standard 3D Point, Standard 3D Scatter, Standard Deviation (Population), Standard Deviation (Sample), Stepwise Regression (Backward), Stepwise Regression (Correlation), Stepwise Regression (Forward), Stepwise Regression (Forward-Backward), Stochastic Processes (Exponential Brownian Motion), Stochastic Processes (Geometric Brownian Motion), Stochastic Processes (Jump Diffusion), Stochastic Processes (Mean Reversion with Jump Diffusion), Stochastic Processes (Mean Reversion), Structural Break, Sum, Time-Series Analysis (Auto), Time-Series Analysis (Double Exponential Smoothing), Time-Series Analysis (Double Moving Average), Time-Series Analysis (Holt-Winter's Additive), Time-Series Analysis (Holt-Winter's Multiplicative), Time-Series Analysis (Seasonal Additive), Time-Series Analysis (Seasonal Multiplicative), Time-Series Analysis (Single Exponential Smoothing), Time-Series Analysis (Single Moving Average), Trend Line (Difference Detrended), Trend Line (Exponential Detrended), Trend Line (Exponential), Trend Line (Linear Detrended), Trend Line (Linear), Trend Line (Logarithmic Detrended), Trend Line (Logarithmic), Trend Line (Moving Average Detrended), Trend Line (Moving Average), Trend Line (Polynomial Detrended), Trend Line (Polynomial), Trend Line (Power Detrended), Trend Line (Power), Trend Line (Rate Detrended), Trend Line (Static Mean Detrended), Trend Line (Static Median Detrended), Variance (Population), Variance (Sample), Volatility: EGARCH, Volatility: EGARCH-T, Volatility: GARCH, Volatility: GARCH-M, Volatility: GJR GARCH, Volatility: GJR TGARCH, Volatility: Log Returns Approach, Volatility: TGARCH, Volatility: TGARCH-M, Yield Curve (Bliss), and Yield Curve (Nelson-Siegel).

2

2. MONTE CARLO RISK SIMULATION

Monte Carlo risk simulation, named for the famous gambling capital of Monaco, is a very potent methodology. For the practitioner, simulation opens the door for solving difficult and complex but practical problems with great ease. Monte Carlo creates artificial futures by generating thousands and even millions of sample paths of outcomes and looks at their prevalent characteristics. For analysts in a company, taking graduate-level advanced math courses is just not logical or practical. A brilliant analyst would use all available tools at his or her disposal to obtain the same answer the easiest and most practical way possible. And in all cases, when modeled correctly, Monte Carlo simulation provides similar answers to the more mathematically elegant methods. So, what is Monte Carlo simulation and how does it work?

2.1 What Is Monte Carlo Simulation?

Monte Carlo simulation in its simplest form is a random number generator that is useful for forecasting, estimation, and risk analysis. A simulation calculates numerous scenarios of a model by repeatedly picking values from a user-predefined probability distribution for the uncertain variables and using those values for the model. As all those scenarios produce associated results in a model, each scenario can have a forecast. Forecasts are events (usually with formulas or functions) that you define as important outputs of the model. These usually are events such as totals, net profit, or gross expenses.

Simplistically, think of the Monte Carlo simulation approach as repeatedly picking golf balls out of a large basket with replacement. The size and shape of the basket depend on the distributional input assumption (e.g., a normal distribution with a mean of 100 and a standard deviation of 10, versus a uniform distribution or a triangular distribution) where some baskets are deeper or more symmetrical than others, allowing certain balls to be pulled out more frequently than others. The number of balls pulled repeatedly depends on the number of trials simulated. For a large model with multiple related assumptions, imagine a very large basket wherein many smaller baskets reside. Each small basket has its own set of golf balls that are bouncing around. Sometimes these small baskets are linked with each other (if there is a correlation between the variables) and the golf balls are bouncing in tandem, while other times the balls are bouncing independently of one another. The balls that are picked each time from these interactions within the model (the large central basket) are tabulated and recorded, providing a forecast output result of the simulation.

2.2 Getting Started with Risk Simulator

2.2.1 A High-Level Overview of the Software

The Risk Simulator software has several different applications including Monte Carlo simulation, forecasting, optimization, and risk analytics.

- The Simulation Module allows you to run simulations in your existing Excel-based models, generate and extract simulation forecasts (distributions of results), perform distributional fitting (automatically finding the best-fitting statistical distribution), compute correlations (maintain relationships among simulated random variables), identify sensitivities (creating tornado and sensitivity charts), test statistical hypotheses (finding statistical differences between pairs of forecasts), run bootstrap simulation (testing the robustness of result statistics), and run custom and nonparametric simulations (simulations using historical data without specifying any distributions or their parameters for forecasting without data or applying expert opinion forecasts).

- The Forecasting Module can be used to generate automatic time-series forecasts (with and without seasonality and trend), multivariate regressions (modeling relationships among variables), nonlinear extrapolations (curve fitting), stochastic processes (random walks, mean-reversions, jump-diffusion, and mixed processes), Box-Jenkins ARIMA (econometric forecasts), Auto ARIMA, basic econometrics and auto econometrics (modeling relationships and generating forecasts), exponential J curves, logistic S curves, GARCH models and their multiple variations (modeling and forecasting volatility), maximum likelihood models for limited dependent variables (logit, tobit, and probit models), Markov chains, trendlines, spline curves, and others.

- The Optimization Module is used for optimizing multiple decision variables subject to constraints to maximize or minimize an objective, and can be run either as a static optimization, dynamic, and stochastic optimization under uncertainty together with Monte Carlo simulation, or as a stochastic optimization with super speed simulations. The software can handle linear and nonlinear optimizations with binary, integer, and continuous variables, as well as generate Markowitz efficient frontiers.

- The Analytical Tools Module allows you to run segmentation clustering, hypothesis testing, statistical tests of raw data, data diagnostics of technical forecasting assumptions (e.g., heteroskedasticity, multicollinearity, and the like), sensitivity and scenario analyses, overlay chart analysis, spider charts, tornado charts, and many other powerful tools.

- ROV BizStats (over 130 business statistics and analytical models).

- ROV Decision Tree (decision tree models, Monte Carlo risk simulation on decision trees, sensitivity analysis, scenario analysis, Bayesian joint and posterior probability updating, expected value of information, MINIMAX, MAXIMIN, risk profiles).

- The Real Options Super Lattice Solver is a software that complements Risk Simulator, used for solving simple to complex real options problems.

The following sections walk you through the basics of the Simulation Module in Risk Simulator, while future chapters provide more details about the applications of other modules. To follow along, make sure you have Risk Simulator installed on your computer to proceed.

In fact, it is highly recommended that you first watch the getting started videos on the web (www.realoptionsvaluation.com/risksimulator.html) or attempt the step-by-step exercises at the end of this chapter before coming back and reviewing the text in this chapter. This approach is recommended because the videos will get you started immediately, as will the exercises, whereas the text in this chapter focuses more on the theory and detailed explanations of the properties of simulation.

2.2.2 Running a Monte Carlo Simulation

Typically, to run a simulation in your existing Excel model, the following steps have to be performed:

1. Start a new simulation profile or open an existing profile.
2. Define input assumptions in the relevant cells.
3. Define output forecasts in the relevant cells.
4. Run simulation.
5. Interpret the results.

If desired, and for practice, open the example file called Basic Simulation Model and follow along with the examples below on creating a simulation. The example file can be found either on the start menu at *Start | Real Options Valuation | Risk Simulator | Examples* or accessed directly through *Risk Simulator | Example Models*.

Starting a New Simulation Profile

To start a new simulation, you will first need to create a simulation profile. A simulation profile contains a complete set of instructions on how you would like to run a simulation. That is, all the assumptions, forecasts, run preferences, and so forth. Having profiles facilitates creating multiple scenarios of simulations. That is, using the same exact model, several profiles can be created, each with its own specific simulation properties and requirements. The same person can create different test scenarios using different distributional assumptions and inputs or multiple persons can test their own assumptions and inputs on the same model.

- Start Excel and create a new model or open an existing one (you can use the Basic Simulation Model example to follow along).
- Click on *Risk Simulator | New Simulation Profile*.
- Specify a title for your simulation as well as all other pertinent information (Figure 2.1).

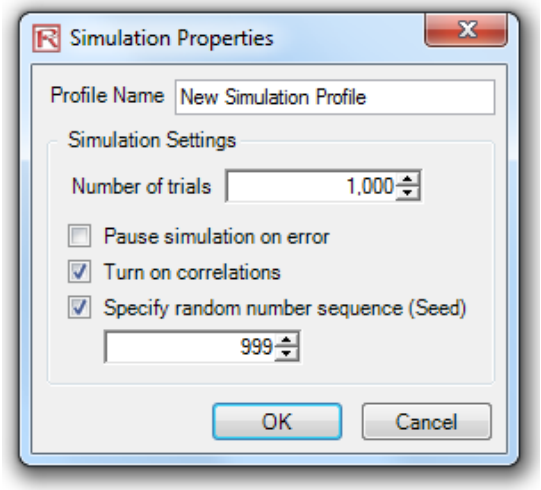


Figure 2.1 – New Simulation Profile

- **Title**: Specifying a simulation title allows you to create multiple simulation profiles in a single Excel model. Thus you can now save different simulation scenario profiles within the same model without having to delete existing assumptions and changing them each time a new simulation scenario is required. You can always change the profile's name later (*Risk Simulator | Edit Profile*).

- **Number of trials**: This is where the number of simulation trials required is entered. That is, running 1,000 trials means that 1,000 different iterations of outcomes based on the input assumptions will be generated. You can change this number as desired, but the input has to be positive integers. The default number of runs is 1,000 trials. You can use precision and error control later in this chapter to automatically help determine how many simulation trials to run (see the section on precision and error control for details).

- **Pause simulation on error**: If checked, the simulation stops every time an error is encountered in the Excel model. That is, if your model encounters a computation error (e.g., some input values generated in a simulation trial may yield a divide by zero error in one of your spreadsheet cells), the simulation stops. This function is important to help audit your model to make sure there are no computational errors in your Excel model. However, if you are sure the model works, then there is no need for this preference to be checked.

- **Turn on correlations**: If checked, correlations between paired input assumptions will be computed. Otherwise, correlations will all be set to zero, and a simulation is run assuming no cross-correlations between input assumptions. As an example, applying correlations will yield more accurate results if, indeed, correlations exist, and will tend to yield a lower forecast confidence if negative correlations exist. After turning on correlations here, you can later set the relevant correlation coefficients on each assumption generated (see the section on correlations for more details).

- **Specify random number sequence**: Simulation by definition will yield slightly different results every time a simulation is run. This characteristic is by virtue of the random number generation routine in Monte Carlo simulation and is a theoretical fact in all random number generators. However, when making

presentations, sometimes you may require the same results (especially when the report being presented shows one set of results and during a live presentation you would like to show the same results being generated, or when you are sharing models with others and would like the same results to be obtained every time), so you would then check this preference and enter in an initial seed number. The seed number can be any positive integer. Using the same initial seed value, the same number of trials, and the same input assumptions, the simulation will always yield the same sequence of random numbers, guaranteeing the same final set of results.

Note that once a new simulation profile has been created, you can come back later and modify these selections. To do so, make sure that the current active profile is the profile you wish to modify, otherwise, click on *Risk Simulator | Change Simulation Profile*, select the profile you wish to change and click *OK* (Figure 2.2 shows an example where there are multiple profiles and how to activate a selected profile). Then, click on *Risk Simulator | Edit Simulation Profile* and make the required changes. You can also duplicate or rename an existing profile. When creating multiple profiles in the same Excel model, make sure to provide each profile a unique name so you can tell them apart later on. Also, these profiles are stored inside hidden sectors of the Excel *.xls file and you do not have to save any additional files. The profiles and their contents (assumptions, forecasts, etc.) are automatically saved when you save the Excel file. Finally, the last profile that is active when you exit and save the Excel file will be the one that is opened the next time the Excel file is accessed.

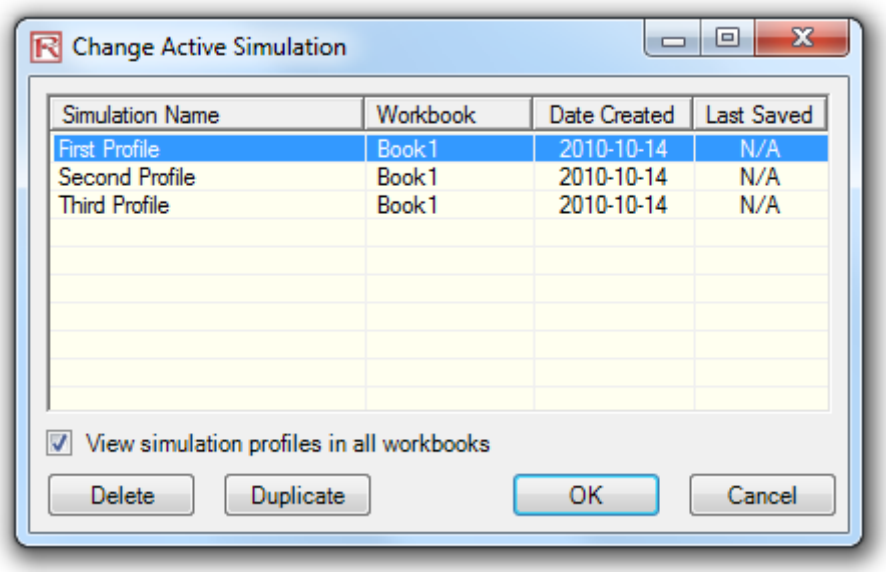


Figure 2.2 – Change Active Simulation

Defining Input Assumptions

The next step is to set input assumptions in your model. Note that assumptions can only be assigned to cells without any equations or functions—typed-in numerical values that are inputs in a model—whereas output forecasts can only be assigned to cells with equations and functions—outputs of a model. Recall that assumptions and forecasts cannot be set unless a simulation profile already exists. Do the following to set new input assumptions in your model:

- Make sure a Simulation Profile exists; open an existing profile or start a new profile (*Risk Simulator | New Simulation Profile*).

- Select the cell you wish to set an assumption on (e.g., cell *G8* in the Basic Simulation Model example).
- Click on *Risk Simulator | Set Input Assumption* or click on the set input assumption icon in the Risk Simulator icon toolbar.
- *Select the relevant distribution* you want, *enter the relevant distribution parameters* (e.g., Triangular distribution with 1, 2, 2.5 as the minimum, most likely, and maximum values), and hit *OK* to insert the input assumption into your model (Figure 2.3).

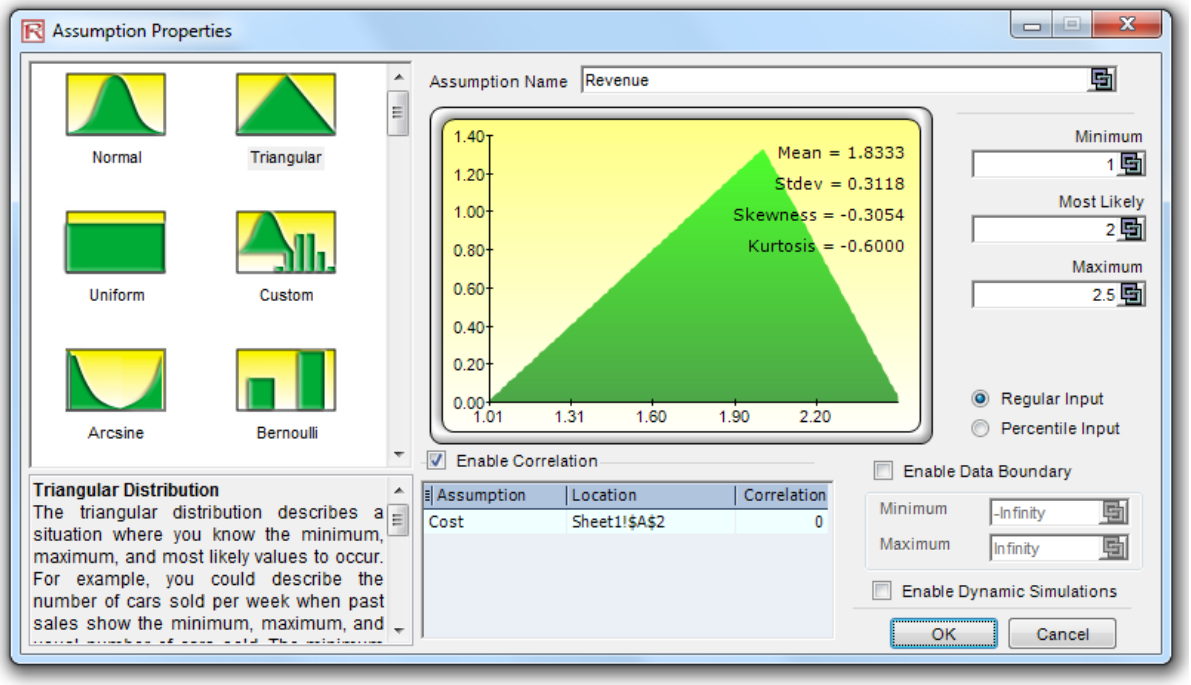


Figure 2.3 – Setting an Input Assumption

Note that you can also set assumptions by selecting the cell you wish to set the assumption on and using the mouse right-click, access the shortcut Risk Simulator menu to set an input assumption. In addition, for expert users, you can set input assumptions using the Risk Simulator RS Functions: select the cell of choice, click on Excel's Insert, Function, select the All Category, and scroll down to the RS functions list (we do not recommend using RS functions unless you are an expert user). For the examples going forward, we suggest following the basic instructions in accessing menus and icons.

As shown in Figure 2.4, there are several key areas in the Assumption Properties worthy of mention.

- **Assumption Name**: This is an optional area to allow you to enter in unique names for the assumptions to help track what each of the assumptions represents. Good modeling practice is to use short but precise assumption names.
- **Distribution Gallery**: This area to the left shows all of the different distributions available in the software. To change the views, right-click anywhere in the gallery

and select large icons, small icons, or list. There are over two dozen distributions available.

- **Input Parameters**: Depending on the distribution selected, the required relevant parameters are shown. You may either enter the parameters directly or link them to specific cells in your worksheet. Hard coding or typing the parameters is useful when the assumption parameters are assumed not to change. Linking to worksheet cells is useful when the input parameters need to be visible or are allowed to be changed (click on the link icon to link an input parameter to a worksheet cell).

- **Enable Data Boundary**: These are typically not used by the average analyst but exist for truncating the distributional assumptions. For instance, if a normal distribution is selected, the theoretical boundaries are between negative infinity and positive infinity. However, in practice, the simulated variable exists only within some smaller range, and this range can then be entered to truncate the distribution appropriately.

- **Correlations**: Pairwise correlations can be assigned to input assumptions here. If correlations are required, remember to check the Turn on Correlations preference by clicking on *Risk Simulator | Edit Simulation Profile*. See the discussion on correlations later in this chapter for more details about assigning correlations and the effects correlations will have on a model. Notice that you can either truncate a distribution or correlate it to another assumption, but not both.

- **Short Descriptions**: These exist for each of the distributions in the gallery. The short descriptions explain when a certain distribution is used as well as the input parameter requirements. See the section in Understanding Probability Distributions for Monte Carlo Simulation for details on each distribution type available in the software.

- **Regular Input and Percentile Input**: This option allows the user to perform a quick due diligence test of the input assumption. For instance, if setting a normal distribution with some mean and standard deviation inputs, you can click on the percentile input to see what the corresponding 10th and 90th percentiles are.

- **Enable Dynamic Simulation**: This option is unchecked by default, but if you wish to run a multidimensional simulation (i.e., if you link the input parameters of the assumption to another cell that is itself an assumption, you are simulating the inputs, or simulating the simulation), then remember to check this option. Dynamic simulation will not work unless the inputs are linked to other changing input assumptions.

Note: If you are following along with the example, continue by setting another assumption on cell G9. This time use the Uniform distribution with a minimum value of 0.9 and a maximum value of 1.1. Then, proceed to defining the output forecasts in the next step.

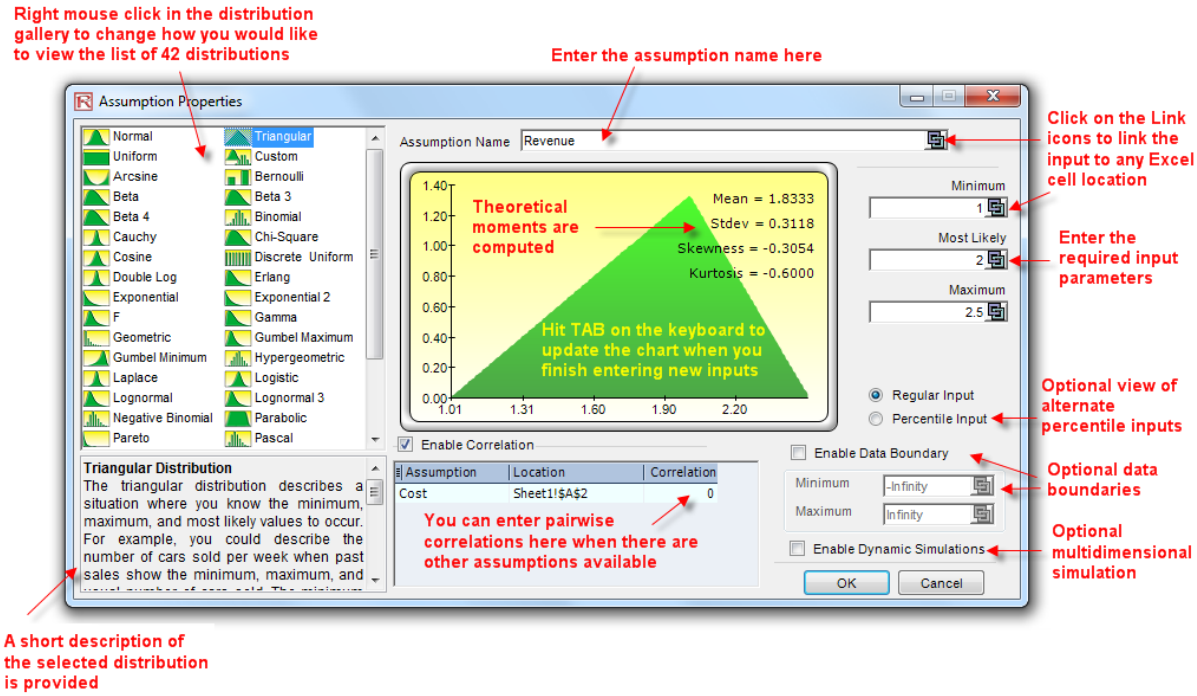


Figure 2.4 – Assumption Properties

Defining Output Forecasts

The next step is to define output forecasts in the model. Forecasts can only be defined on output cells with equations or functions. The following describes the set forecast process:

- Select the cell you wish to set a forecast (e.g., cell *G10* in the Basic Simulation Model example).
- Click on *Risk Simulator* | *Set Output Forecast* or click on the set output forecast icon on the Risk Simulator icon toolbar (Figure 1.3).
- Enter the relevant information and click *OK*.

Note that you can also set output forecasts by selecting the cell you wish to set the forecast on and using the mouse right-click, access the shortcut Risk Simulator menu to set an output forecast. Figure 2.5 illustrates the set forecast properties.

- **Forecast Name:** Specify the name of the forecast cell. This is important because when you have a large model with multiple forecast cells, naming the forecast cells individually allows you to access the right results quickly. Do not underestimate the importance of this simple step. Good modeling practice is to use short but precise forecast names.
- **Forecast Precision**: Instead of relying on a guesstimate of how many trials to run in your simulation, you can set up precision and error controls. When an error-precision combination has been achieved in the simulation, the simulation will pause and inform you of the precision achieved, making the required number of simulation trials an automated process rather than a guessing game. Review the section on error and precision control later in this chapter for more specific details.

- **Show Forecast Window**: Allows the user to show or not show a particular forecast window. The default is to always show a forecast chart.

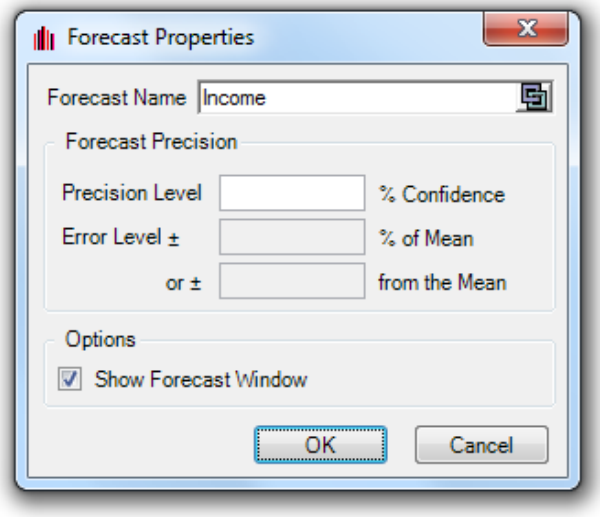


Figure 2.5 – Set Output Forecast

Running the Simulation

If everything looks right, simply click on *Risk Simulator | Run Simulation* or click on the *Run icon* on the Risk Simulator toolbar and the simulation will proceed. You may also reset a simulation after it has run to rerun it (*Risk Simulator | Reset Simulation* or the reset simulation icon on the toolbar) or to pause it during a run. Also, the step function (*Risk Simulator | Step Simulation* or the *step simulation icon* on the toolbar) allows you to simulate a single trial, one at a time, useful for educating others on simulation (i.e., you can show that at each trial, all the values in the assumption cells are being replaced and the entire model is recalculated each time). You can also access the run simulation menu by right-clicking anywhere in the model and selecting Run Simulation.

Risk Simulator also allows you to run the simulation at extremely fast speed, called Super Speed. To do this, click on *Risk Simulator | Run Super Speed Simulation* or use the run super speed icon. Notice how much faster the super speed simulation runs. In fact, for practice, *Reset Simulation* and then *Edit Simulation Profile* and change the *Number of Trials* to 100,000, and Run Super Speed. It should only take a few seconds to run. However, please be aware that super speed simulation will not run if the model has errors, VBA (visual basic for applications), or links to external data sources or applications. In such situations, you will be notified and the regular speed simulation will be run instead. Regular speed simulations are always able to run even with errors, VBA, or external links.

Interpreting the Forecast Results

The final step in Monte Carlo simulation is to interpret the resulting forecast charts. Figures 2.6 through 2.13 show the forecast chart and the corresponding statistics generated after running the simulation. Typically, the following elements are important in interpreting the results of a simulation:

Forecast Chart

The forecast chart shown in Figure 2.6 is a probability histogram that shows the frequency counts of values occurring in the total number of trials simulated. The vertical bars show the frequency of a particular x value occurring out of the total number of trials, while the

cumulative frequency (smooth line) shows the total probabilities of all values at and below x occurring in the forecast.

Forecast Statistics

The forecast statistics shown in Figure 2.7 summarize the distribution of the forecast values in terms of the four moments of a distribution. See the Understanding the Forecast Statistics section later in this chapter for more details on what some of these statistics mean. You can rotate between the histogram and statistics tabs by depressing the space bar.

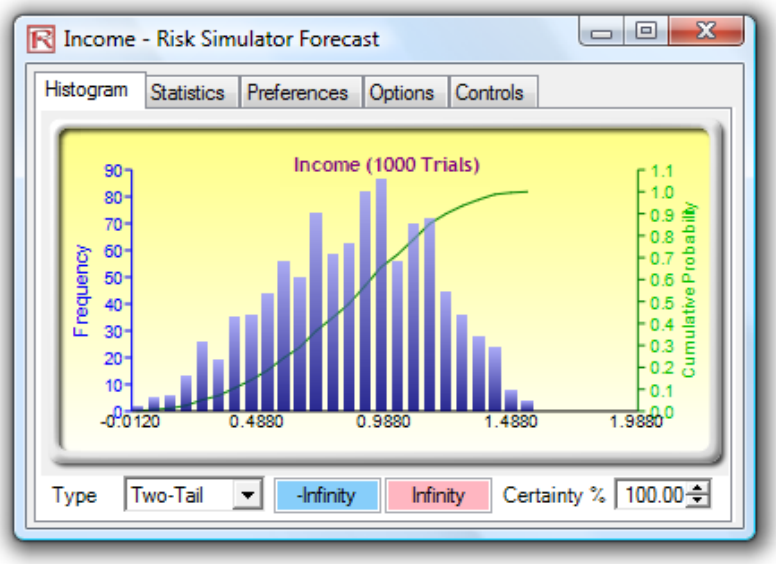
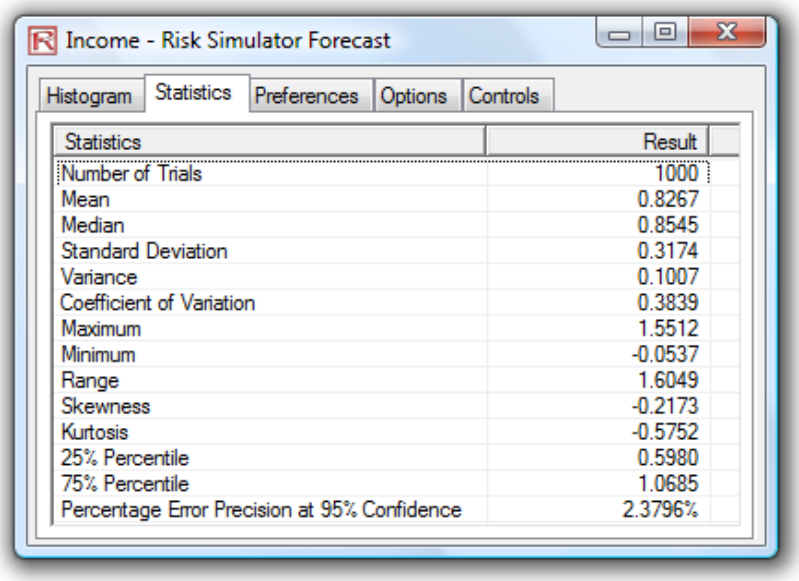


Figure 2.6 – Forecast Chart



Statistics	Result
Number of Trials	1000
Mean	0.8267
Median	0.8545
Standard Deviation	0.3174
Variance	0.1007
Coefficient of Variation	0.3839
Maximum	1.5512
Minimum	-0.0537
Range	1.6049
Skewness	-0.2173
Kurtosis	-0.5752
25% Percentile	0.5980
75% Percentile	1.0685
Percentage Error Precision at 95% Confidence	2.3796%

Figure 2.7 – Forecast Statistics

2.2.3 Forecast Chart Tabs

Preferences

The preferences tab in the forecast chart (Figure 2.8A) allows you to change the look and feel of the charts. For instance, if *Always On Top* is selected, the forecast charts will always be visible regardless of what other software are running on your computer. *Histogram Resolution* allows you to change the number of bins of the histogram, anywhere from 5 bins to 100 bins. Also, the *Data Update* feature allows you to control how fast the simulation runs versus how often the forecast chart is updated. For example, viewing the forecast chart updated at almost every trial will slow down the simulation as more memory is being allocated to updating the chart versus running the simulation. This is merely a user preference and in no way changes the results of the simulation, just the speed of completing the simulation. To further increase the speed of the simulation, you can minimize Excel while the simulation is running, thereby reducing the memory required to visibly update the Excel spreadsheet and freeing up the memory to run the simulation. The *Clear All* and *Minimize All* controls all the open forecast charts.

Options

As shown in Figure 2.8B, this forecast chart feature allows you to show all the forecast data or to filter in/out values that fall within either some specified interval or some standard deviation you choose. Also, the precision level can be set here for this specific forecast to show the error levels in the statistics view. See the section on error and precision control later in this chapter for more details. Show the following statistic on histogram is a user preference for whether the mean, median, first quartile, and fourth quartile lines (25th and 75th percentiles) should be displayed on the forecast chart.

Controls

As shown in Figure 2.8C, this tab has all the functionalities in allowing you to change the type, color, size, zoom, tilt, 3D, and other things in the forecast chart, as well as to generate overlay charts (PDF, CDF) and run distributional fitting on your forecast data (see the Data Fitting sections for more details on this methodology).

Global View versus Normal View

Figures 2.8A to 2.8C show the forecast chart's Normal View where the forecast chart user interface is divided into tabs, making it small and compact. In contrast, Figure 2.9 shows the Global View where all elements are located in a single interface. The results are identical in both views and selecting which view is a matter of personal preference. You can switch between these two views by clicking on the link, located at the top right corner, called *Global View* and *Local View*.

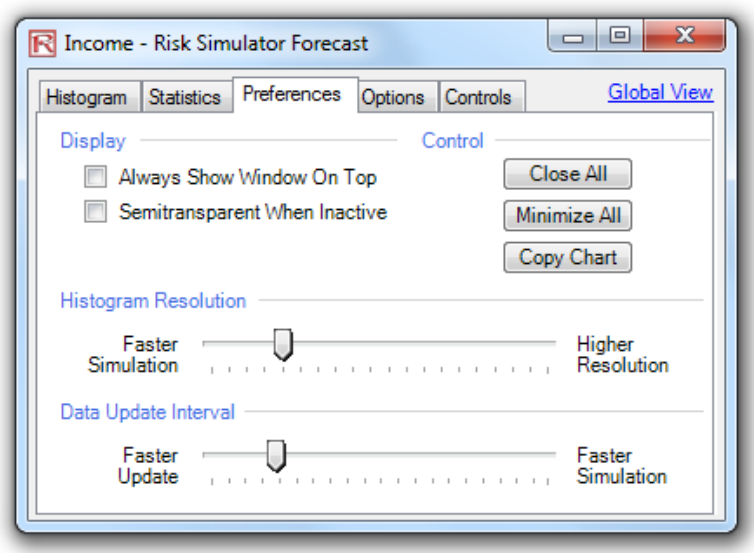


Figure 2.8A – Forecast Chart Preferences

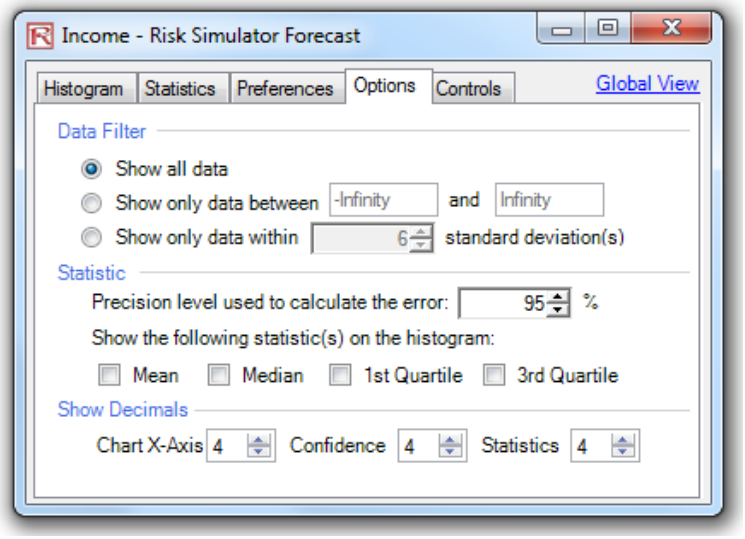


Figure 2.8B – Forecast Chart Options

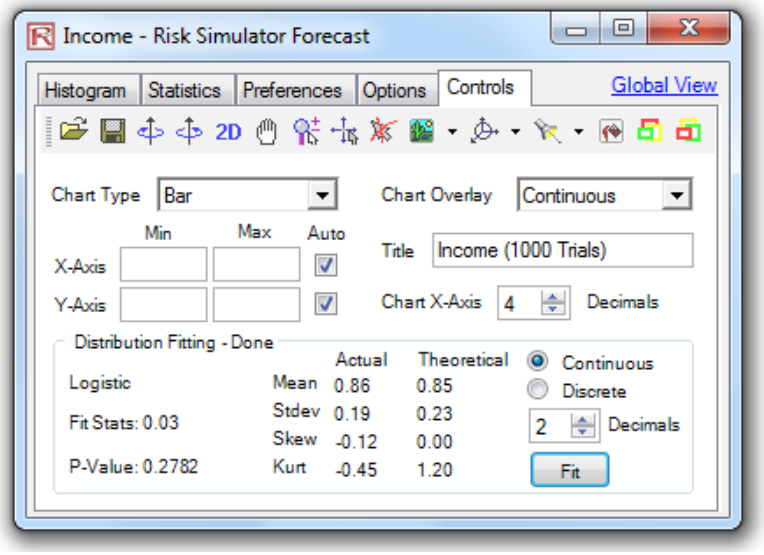


Figure 2.8C – Forecast Chart Controls

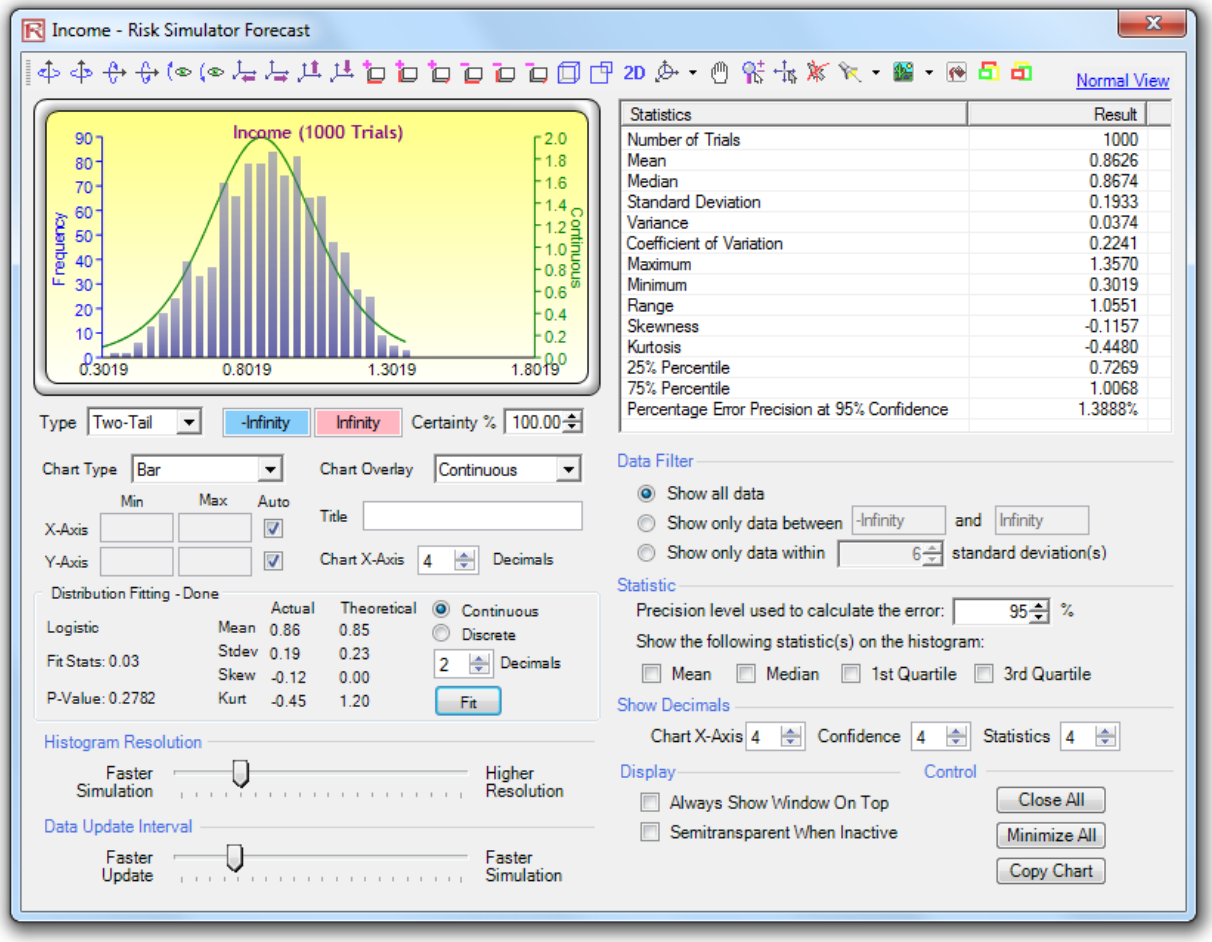


Figure 2.9 – Forecast Chart Global View

2.3.4 Using Forecast Charts and Confidence Intervals

In forecast charts, you can determine the probability of occurrence called confidence intervals. That is, given two values, what are the chances that the outcome will fall between these two values? Figure 2.10 illustrates that there is a 90% probability that the final outcome (in this case, the level of income) will be between $0.2653 and $1.3230. The two-tailed confidence interval can be obtained by first selecting *Two-Tail* as the type, entering the desired certainty value (e.g., *90*) and hitting *TAB* on the keyboard. The two computed values corresponding to the certainty value will then be displayed. In this example, there is a 5% probability that income will be below $0.2653 and another 5% probability that income will be above $1.3230. That is, the two-tailed confidence interval is a symmetrical interval centered on the median, or 50th percentile, value. Thus, both tails will have the same probability.

Alternatively, a one-tail probability can be computed. Figure 2.11 shows a left-tail selection at 95% confidence (i.e., choose *Left-Tail* ≤ as the type, enter *95* as the certainty level, and hit *TAB* on the keyboard). This means that there is a 95% probability that the income will be below $1.3230 or a 5% probability that income will be above $1.3230, corresponding perfectly with the results seen in Figure 2.10.

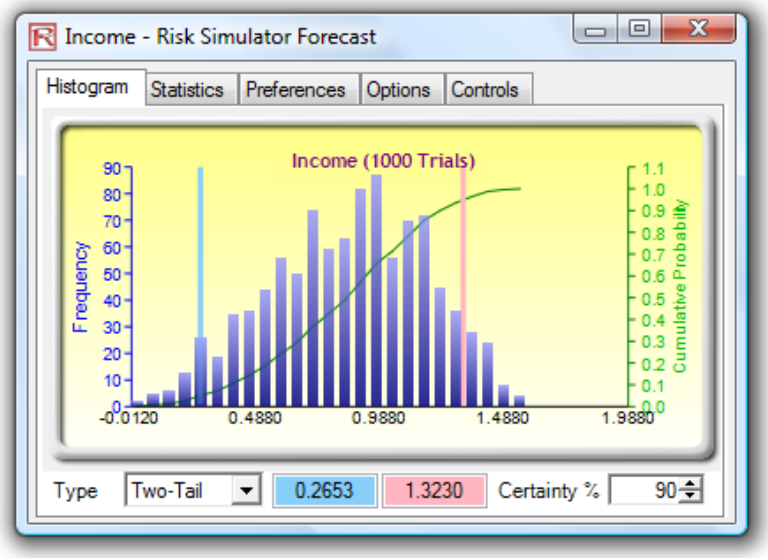


Figure 2.10 – Forecast Chart Two-Tail Confidence Interval

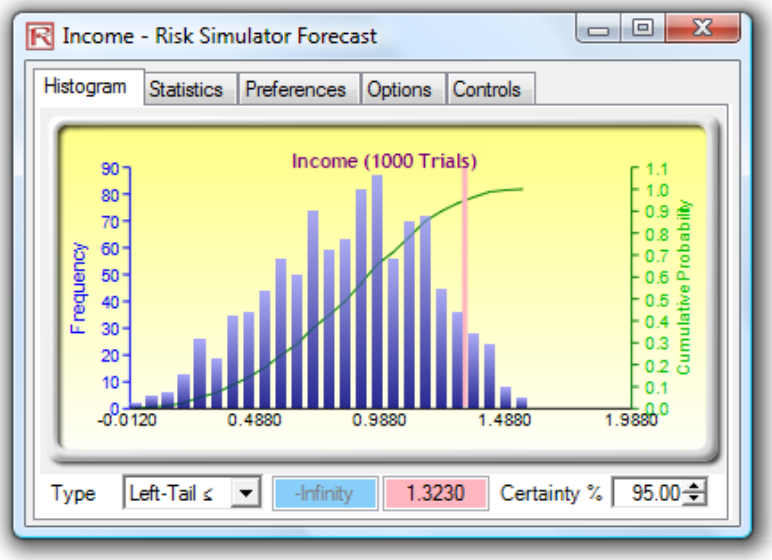


Figure 2.11 – Forecast Chart One-Tail Confidence Interval

In addition to evaluating what the confidence interval is (i.e., given a probability level and finding the relevant income values), you can determine the probability of a given income value. For instance, what is the probability that income will be less than or equal to $1? To obtain the answer, select the *Left-Tail* ≤ probability type, enter *1* into the value input box, and hit TAB. The corresponding certainty will then be computed (in this case, as shown in Figure 2.12, there is a 67.70% probability income will be at or below $1).

For the sake of completeness, you can select the *Right-Tail* > probability type, and enter the value *1* in the value input box, and hit TAB. The resulting probability indicates the right-tail probability past the value 1, that is, the probability of income exceeding $1 (in this case, as shown in Figure 2.13, we see that there is a 32.30% probability of income exceeding $1). The sum of 67.70% and 32.30% is, of course, 100%, the total probability under the curve.

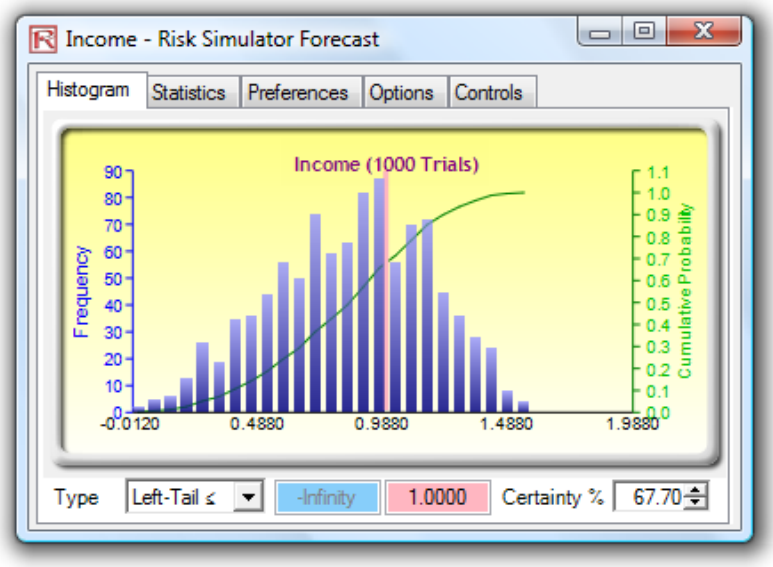


Figure 2.12 – Forecast Chart Probability Evaluation

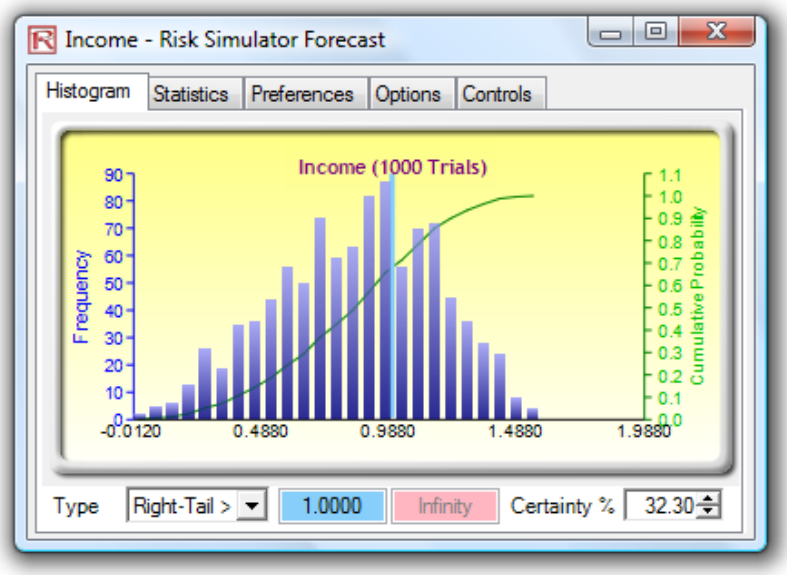


Figure 2.13 – Forecast Chart Probability Evaluation

TIPS

- The forecast window is resizable by clicking on and dragging the bottom right corner of the forecast window.
- It is also advisable that the current simulation be reset (*Risk Simulator | Reset Simulation*) before rerunning a simulation.
- Remember that you will need to hit TAB on the keyboard to update the chart and results when you type in the certainty values or right- and left-tail values.

- You can also hit the spacebar on the keyboard repeatedly to cycle among the histogram to statistics, preferences, options, and control tabs.
- In addition, if you click on *Risk Simulator | Options* you can access several different options for Risk Simulator, including allowing Risk Simulator to start each time Excel starts or to only start when you want it to (by going to *Start | Programs | Real Options Valuation | Risk Simulator | Risk Simulator*), changing the cell colors of assumptions and forecasts, and turning cell comments on and off (cell comments will allow you to see which cells are input assumptions and which are output forecasts as well as their respective input parameters and names). Do spend some time playing around with the forecast chart outputs and various bells and whistles, especially the Controls tab.

2.3 Correlations and Precision Control

2.3.1 The Basics of Correlations

The correlation coefficient is a measure of the strength and direction of the relationship between two variables, and it can take on any value between –1.0 and +1.0. That is, the correlation coefficient can be decomposed into its sign (positive or negative relationship between two variables) and the magnitude or strength of the relationship (the higher the absolute value of the correlation coefficient, the stronger the relationship).

The correlation coefficient can be computed in several ways. The first approach is to manually compute the correlation, r, of two variables, x and y, using:

$$r_{x,y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

The second approach is to use Excel's CORREL function. For instance, if the 10 data points for x and y are listed in cells A1:B10, then the Excel function to use is CORREL (A1:A10, B1:B10).

The third approach is to run Risk Simulator's Multi-Fit Tool, and the resulting correlation matrix will be computed and displayed.

It is important to note that correlation does not imply causation. Two completely unrelated random variables might display some correlation but this does not imply any causation between the two (e.g., sunspot activity and events in the stock market are correlated but there is no causation between the two).

There are two general types of correlations: parametric and nonparametric correlations. Pearson's correlation coefficient is the most common correlation measure and is usually referred to simply as the correlation coefficient. However, Pearson's correlation is a parametric measure, which means that it requires both correlated variables to have an underlying normal distribution and that the relationship between the variables is linear. When these conditions are violated, which is often the case in Monte Carlo simulation, the nonparametric counterparts become more important. Spearman's rank correlation and Kendall's tau are the two alternatives. The Spearman correlation is most commonly used and is most appropriate when applied in the context of Monte Carlo simulation—there is no dependence on normal distributions or linearity, meaning that correlations between different variables with different

distribution can be applied. To compute the Spearman correlation, first rank all the x and y variable values and then apply the Pearson's correlation computation.

In the case of Risk Simulator, the correlation used is the more robust nonparametric Spearman's rank correlation. However, to simplify the simulation process, and to be consistent with Excel's correlation function, the correlation inputs required are the Pearson's correlation coefficient. Risk Simulator will then apply its own algorithms to convert them into Spearman's rank correlation, thereby simplifying the process. However, to simplify the user interface, we allow users to enter the more common Pearson's product-moment correlation (e.g., computed using Excel's CORREL function), while in the mathematical codes, we convert these simple correlations into Spearman's rank-based correlations for distributional simulations.

2.3.2 Applying Correlations in Risk Simulator

Correlations can be applied in Risk Simulator in several ways:

- When defining assumptions (*Risk Simulator | Set Input Assumption*), simply *enter the correlations* into the correlation matrix grid in the *Distribution Gallery*.
- With existing data, run the Multi-Fit tool (*Risk Simulator | Tools | Distributional Fitting | Multiple Variables*) to perform distributional fitting and to obtain the correlation matrix between pairwise variables. If a simulation profile exists, the assumptions fitted will automatically contain the relevant correlation values.
- With existing assumptions, you can click on *Risk Simulator | Tools | Edit Correlations* to enter the pairwise correlations of all the assumptions directly in one user interface.

Note that the correlation matrix must be positive definite. That is, the correlation must be mathematically valid. For instance, suppose you are trying to correlate three variables: grades of graduate students in a particular year, the number of beers they consume a week, and the number of hours they study a week. One would assume that the following correlation relationships exist:

Grades and Beer:	–	The more they drink, the lower the grades (no-show on exams)
Grades and Study:	+	The more they study, the higher the grades
Beer and Study:	–	The more they drink, the less they study (drunk and partying)

However, if you input a negative correlation between Grades and Study, and assuming that the correlation coefficients have high magnitudes, the correlation matrix will be nonpositive definite. It would defy logic, correlation requirements, and matrix mathematics. However, smaller coefficients can sometimes still work even with the bad logic. When a nonpositive or bad correlation matrix is entered, Risk Simulator will automatically inform you, and offers to adjust these correlations to something that is semipositive definite while still maintaining the overall structure of the correlation relationship (the same signs as well as the same relative strengths).

2.3.3 The Effects of Correlations in Monte Carlo Simulation

Although the computations required to correlate variables in a simulation are complex, the resulting effects are fairly clear. Figure 2.14 shows a simple correlation model (Correlation Effects Model in the example folder). The calculation for revenue is simply price multiplied by quantity. The same model is replicated for no correlations, positive correlation (+0.8), and negative correlation (–0.8) between price and quantity.

Correlation Model

	Without Correlation	Positive Correlation	Negative Correlation
Price	$2.00	$2.00	$2.00
Quantity	1.00	1.00	1.00
Revenue	$2.00	$2.00	$2.00

Figure 2.14 – Simple Correlation Model

The resulting statistics are shown in Figure 2.15. Notice that the standard deviation of the model without correlations is 0.1450, compared to 0.1886 for the positive correlation and 0.0717 for the negative correlation. That is, for simple models, negative correlations tend to reduce the average spread of the distribution and create a tight and more concentrated forecast distribution as compared to positive correlations with larger average spreads. However, the mean remains relatively stable. This implies that correlations do little to change the expected value of projects but can reduce or increase a project's risk.

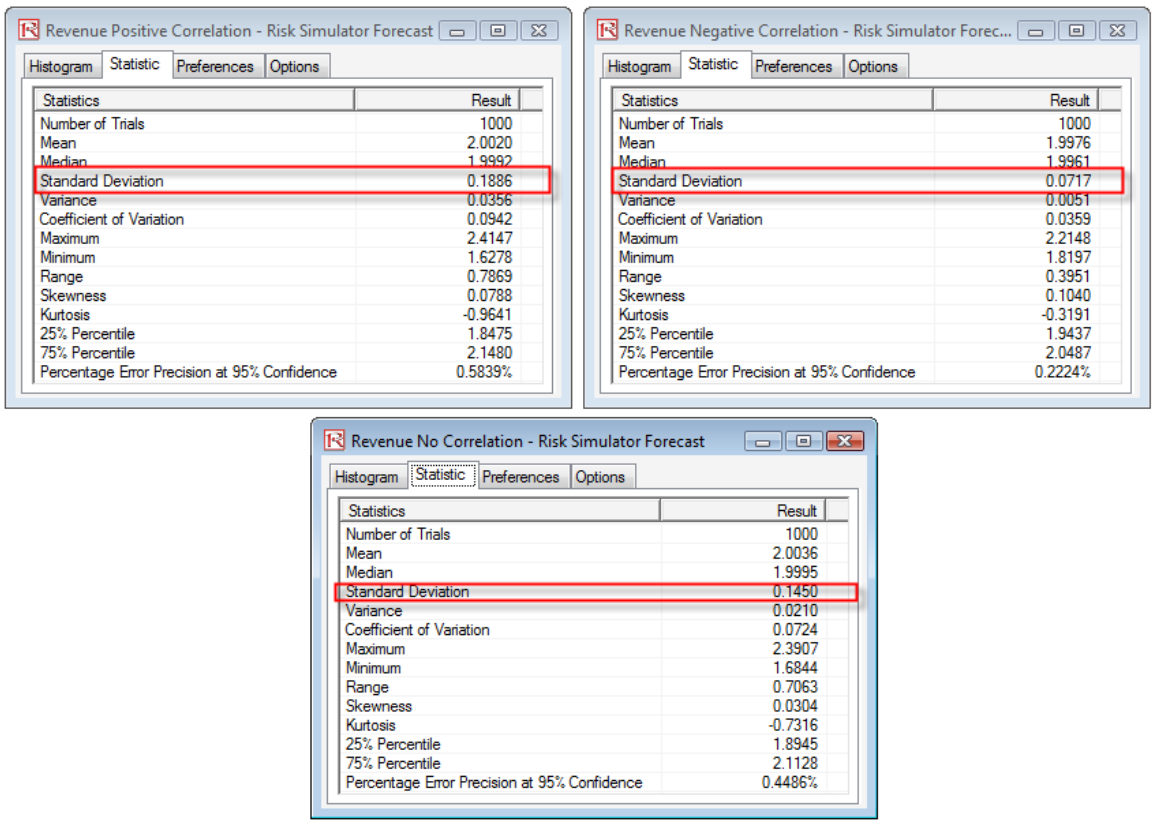


Figure 2.15 – Correlation Results

Figure 2.16 illustrates the results after running a simulation, extracting the raw data of the assumptions and computing the correlations between the variables. The figure shows that the input assumptions are recovered in the simulation. That is, you enter +0.8 and –0.8 correlations and the resulting simulated values have the same correlations.

Price Positive Correlation	Quantity Positive Correlation		Price Negative Correlation	Quantity Negative Correlation	
1.95	0.91		1.89	1.06	
1.92	0.95		1.98	1.05	
2.02	1.04	Pearson's Correlation:	1.89	1.09	Pearson's Correlation:
2.04	1.03		1.88	1.04	
1.89	0.91	**0.80**	1.96	0.93	**-0.80**
1.98	1.05		2.02	0.93	
2.05	1.03		2.00	1.02	
1.87	0.91		1.86	1.04	
1.84	0.91		1.96	1.02	
2.06	1.03		1.90	1.02	
1.98	1.01		1.92	1.10	

Figure 2.16 – Correlations Recovered

2.3.4 Precision and Error Control

One very powerful tool in Monte Carlo simulation is that of precision control. For instance, how many trials are considered sufficient to run in a complex model? Precision control takes the guesswork out of estimating the relevant number of trials by allowing the simulation to stop if the level of prespecified precision is reached.

The precision control functionality lets you set how precise you want your forecast to be. Generally speaking, as more trials are calculated, the confidence interval narrows and the statistics become more accurate. The precision control feature in Risk Simulator uses the characteristic of confidence intervals to determine when a specified accuracy of a statistic has been reached. For each forecast, you can set the specific confidence interval for the precision level.

Make sure that you do not confuse three very different terms: error, precision, and confidence. Although they sound similar, the concepts are significantly different from one another. A simple illustration is in order. Suppose you are a taco shell manufacturer and are interested in finding out how many broken taco shells there are on average in a box of 100 shells. One way to do this is to collect a sample of prepackaged boxes of 100 taco shells, open them, and count how many of them are actually broken. You manufacture 1 million boxes a day (this is your population) but you randomly open only 10 boxes (this is your sample size, also known as your number of trials in a simulation). The number of broken shells in each box is as follows: 24, 22, 4, 15, 33, 32, 4, 1, 45, and 2. The calculated average number of broken shells is 18.2. Based on these 10 samples or trials, the average is 18.2 units, while based on the sample, the 80% confidence interval is between 2 and 33 units (that is, 80% of the time, the number of broken shells is between 2 and 33 based on this sample size or number of trials run). However, how sure are you that 18.2 is the correct average? Are 10 trials sufficient to establish this? The confidence interval between 2 and 33 is too wide and too variable. Suppose you require a more accurate average value where the error is ±2 taco shells 90% of the time—this means that if you open all 1 million boxes manufactured in a day, 900,000 of these boxes will have broken taco shells on average at some mean unit ±2 taco shells. How many more taco shell boxes would you then need to sample (or trials run) to obtain this level of precision? Here, the 2 taco shells is the error level while the 90% is the level of precision. If sufficient numbers of trials are run, then the 90% confidence interval will be identical to the 90% precision level, where a more precise measure of the average is obtained such that 90% of the time, the error and, hence, the confidence will be ±2 taco shells. As an example, say the average is 20 units, then the 90%

confidence interval will be between 18 and 22 units with this interval being precise 90% of the time, where in opening all 1 million boxes, 900,000 of them will have between 18 and 22 broken taco shells. The number of trials required to hit this precision is based on the sampling error equation of $\bar{x} \pm Z\frac{s}{\sqrt{n}}$, where $Z\frac{s}{\sqrt{n}}$ is the error of 2 taco shells, $\bar{x}$ is the sample average, Z is the standard-normal Z-score obtained from the 90% precision level, s is the sample standard deviation, and n is the number of trials required to hit this level of error with the specified precision. Figures 2.17 and 2.18 illustrate how precision control can be performed on multiple simulated forecasts in Risk Simulator. This feature prevents the user from having to decide how many trials to run in a simulation and eliminates all possibilities of guesswork. Figure 2.17 illustrates the forecast chart with a 95% precision level set. This value can be changed and will be reflected in the Statistics tab as shown in Figure 2.18.

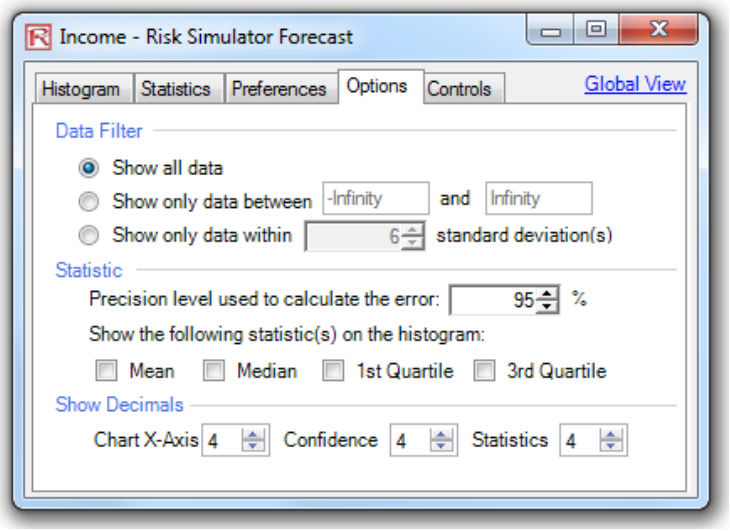
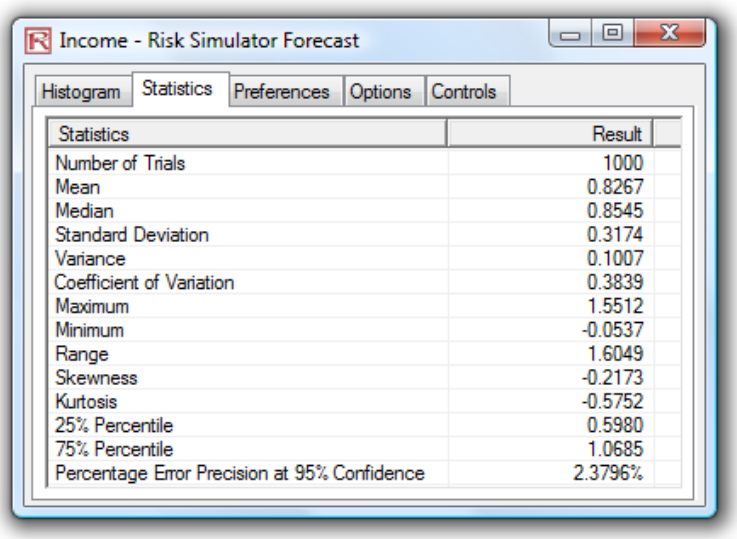


Figure 2.17 – Setting the Forecast's Precision Level



Statistics	Result
Number of Trials	1000
Mean	0.8267
Median	0.8545
Standard Deviation	0.3174
Variance	0.1007
Coefficient of Variation	0.3839
Maximum	1.5512
Minimum	-0.0537
Range	1.6049
Skewness	-0.2173
Kurtosis	-0.5752
25% Percentile	0.5980
75% Percentile	1.0685
Percentage Error Precision at 95% Confidence	2.3796%

Figure 2.18 – Computing the Error

2.3.5 Understanding the Forecast Statistics

Most distributions can be defined up to four moments. The first moment describes a distribution's location or central tendency (expected returns); the second moment describes its width or spread (risks); the third moment, its directional skew (most probable events); and the fourth moment, its peakedness or thickness in the tails (catastrophic losses or gains). All four moments should be calculated in practice and interpreted to provide a more comprehensive view of the project under analysis. Risk Simulator provides the results of all four moments in its *Statistics* view in the forecast charts.

Measuring the Center of the Distribution—the First Moment

The first moment of a distribution measures the expected rate of return on a particular project. It measures the location of the project's scenarios and possible outcomes on average. The common statistics for the first moment include the mean (average), median (center of a distribution), and mode (most commonly occurring value). Figure 2.19 illustrates the first moment—where, in this case, the first moment of this distribution is measured by the mean *(μ),* or average, value.

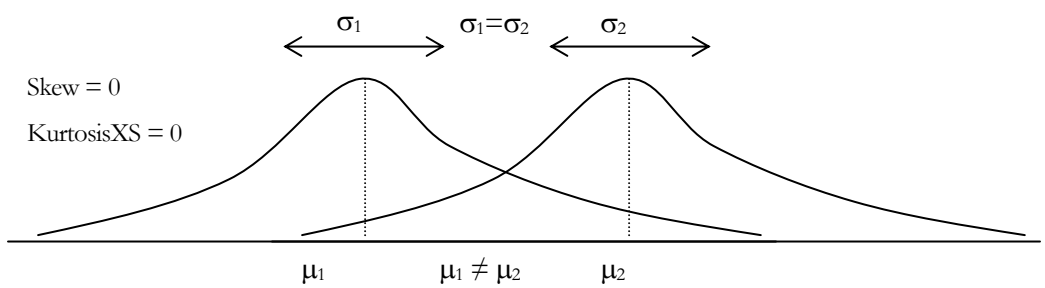


Figure 2.19 – First Moment

Measuring the Spread of the Distribution—the Second Moment

The second moment measures the spread of a distribution, which is a measure of risk. The spread, or width, of a distribution measures the variability of a variable, that is, the potential that the variable can fall into different regions of the distribution—in other words, the potential scenarios of outcomes. Figure 2.20 illustrates two distributions with identical first moments (identical means) but very different second moments or risks. The visualization becomes clearer in Figure 2.21. As an example, suppose there are two stocks and the first stock's movements (illustrated by the darker line) with the smaller fluctuation is compared against the second stock's movements (illustrated by the dotted line) with a much higher price fluctuation. Clearly an investor would view the stock with the wilder fluctuation as riskier because the outcomes of the more risky stock are relatively more unknown than the less risky stock. The vertical axis in Figure 2.21 measures the stock prices, thus, the more risky stock has a wider range of potential outcomes. This range is translated into a distribution's width (the horizontal axis) in Figure 2.20, where the wider distribution represents the riskier asset. Hence, width, or spread, of a distribution measures a variable's risks.

Notice that in Figure 2.20, both distributions have identical first moments, or central tendencies, but the distributions are clearly very different. This difference in the distributional width is measurable. Mathematically and statistically, the width, or risk, of a variable can be measured through several different statistics, including the range, standard deviation (σ), variance, coefficient of variation, and percentiles.

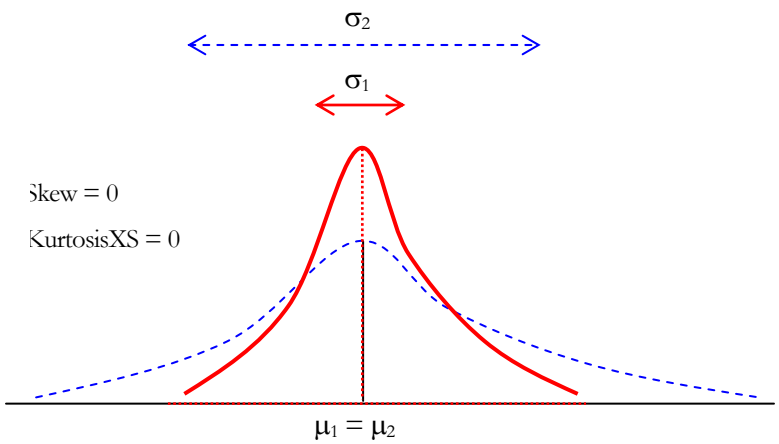


Figure 2.20 – Second Moment

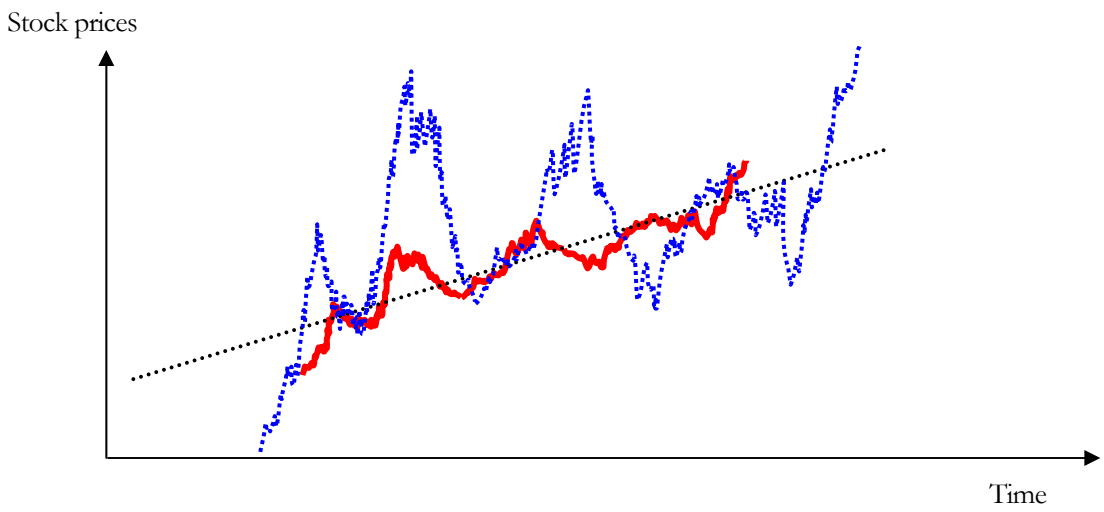


Figure 2.21 – Stock Price Fluctuations

Measuring the Skew of the Distribution—the Third Moment

The third moment measures a distribution's skewness, that is, how the distribution is pulled to one side or the other. Figure 2.22 illustrates a negative-skew, or left-skew, where the tail of the distribution points to the left. Figure 2.23 illustrates a positive-skew or right-skew, where the tail of the distribution points to the right. The mean is always skewed toward the tail of the distribution, while the median remains constant. Another way of seeing this relationship is that the mean moves but the standard deviation, variance, or width may still remain constant. If the third moment is not considered, then looking only at the expected returns (e.g., median or mean) and risk (standard deviation), a positively skewed project might be incorrectly chosen! For example, if the horizontal axis represents the net revenues of a project, then clearly a left, or negatively, skewed distribution might be preferred because there is a higher probability of greater returns (Figure 2.22) as compared to a higher probability for lower level returns (Figure

2.23). Thus, in a skewed distribution, the median is a better measure of returns, as the medians for both Figures 2.22 and 2.23 are identical, risks are identical, and, hence, a project with a negatively skewed distribution of net profits is a better choice. Failure to account for a project's distributional skewness may mean that the incorrect project could be chosen (e.g., two projects may have identical first and second moments, that is, they both have identical returns and risk profiles, but their distributional skews may be very different).

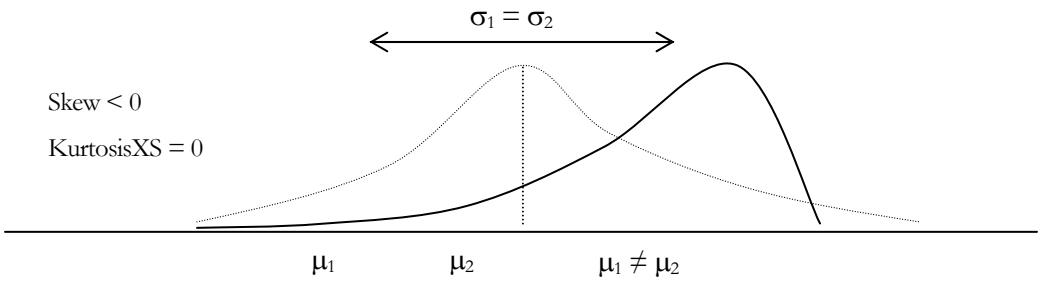


Figure 2.22 – Third Moment (Left Skew)

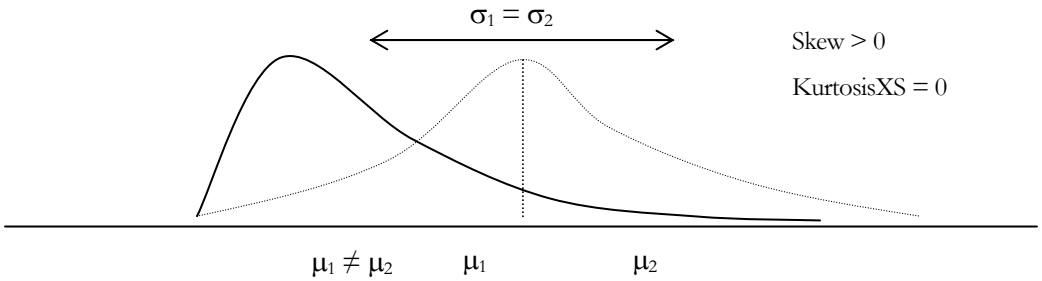


Figure 2.23 – Third Moment (Right Skew)

Measuring the Catastrophic Tail Events in a Distribution—the Fourth Moment

The fourth moment, or kurtosis, measures the peakedness of a distribution. Figure 2.24 illustrates this effect. The background (denoted by the dotted line) is a normal distribution with a kurtosis of 3.0, or an excess kurtosis (KurtosisXS) of 0.0. Risk Simulator's results show the KurtosisXS value, using 0 as the normal level of kurtosis, which means that a negative KurtosisXS indicates flatter tails (platykurtic distributions like the uniform distribution), while positive values indicate fatter tails (leptokurtic distributions like the student's t or lognormal distributions). The distribution depicted by the bold line has a higher excess kurtosis, thus the area under the curve is thicker at the tails with less area in the central body. This condition has major impacts on risk analysis. As shown for the two distributions in Figure 2.24, the first three moments (mean, standard deviation, and skewness) can be identical, but the fourth moment (kurtosis) is different. This condition means that, although the returns and risks are identical, the probabilities of extreme and catastrophic events (potential large losses or large gains) occurring are higher for a high kurtosis distribution (e.g., stock market returns are leptokurtic, or have high kurtosis). Ignoring a project's kurtosis may be detrimental. Typically, a higher excess kurtosis value indicates that the downside risks are higher (e.g., the Value at Risk of a project might be significant).

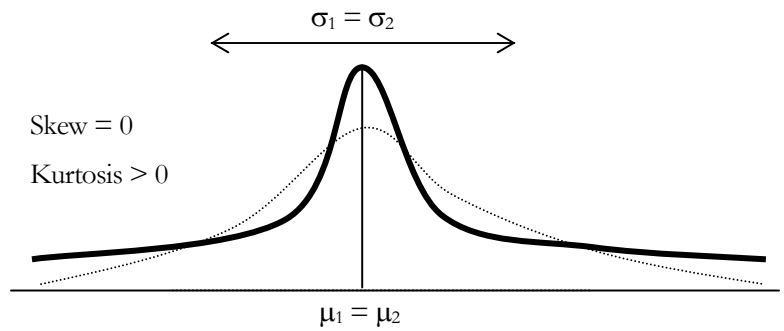


Figure 2.24 – Fourth Moment

The Functions of Moments

Ever wonder why these risk statistics are called "moments"? In mathematical vernacular, *moment* means raised to the power of some value. In other words, the third moment implies that in an equation, three is most probably the highest power. In fact, the equations below illustrate the mathematical functions and applications of some moments for a sample statistic. For example, notice that the highest power for the first moment average is one, the second moment standard deviation is two, the third moment skew is three, and the highest power for the fourth moment is four.

First Moment: Arithmetic Average or Simple Mean (Sample)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The Excel equivalent function is AVERAGE.

Second Moment: Standard Deviation (Sample)

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

The Excel equivalent function is STDEV for a sample standard deviation.

The Excel equivalent function is STDEVP for a population standard deviation.

Third Moment: Skew (Sample)

$$skew = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \frac{(x_i - \bar{x})^3}{s}$$

The Excel equivalent function is SKEW.

Fourth Moment: Kurtosis (Sample)

$$kurtosis = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \frac{(x_i - \bar{x})^4}{s} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

The Excel equivalent function is KURT.

2.3.6 Understanding Probability Distributions for Monte Carlo Simulation

This section demonstrates the power of Monte Carlo simulation, but to get started with simulation, one first needs to understand the concept of *probability distributions*. To begin to understand probability, consider this example: You want to look at the distribution of nonexempt wages within one department of a large company. First, you gather raw data—in this case, the wages of each nonexempt employee in the department. Second, you organize the data into a meaningful format and plot the data as a frequency distribution on a chart. To create a frequency distribution, you divide the wages into group intervals and list these intervals on the chart's horizontal axis. Then you list the number or frequency of employees in each interval on the chart's vertical axis. Now you can easily see the distribution of nonexempt wages within the department.

A glance at the chart illustrated in Figure 2.25 reveals that most of the employees (approximately 60 out of a total of 180) earn from $7.00 to $9.00 per hour.

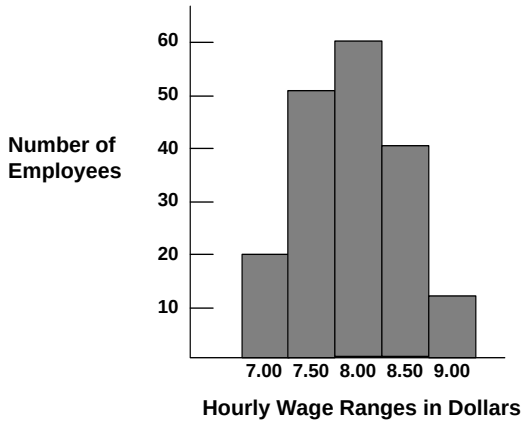


Figure 2.25 – Frequency Histogram I

You can chart this data as a probability distribution. A probability distribution shows the number of employees in each interval as a fraction of the total number of employees. To create a probability distribution, you divide the number of employees in each interval by the total number of employees and list the results on the chart's vertical axis.

The chart in Figure 2.26 shows you the number of employees in each wage group as a fraction of all employees; you can estimate the likelihood or probability that an employee drawn at random from the whole group earns a wage within a given interval. For example, assuming the same conditions exist at the time the sample was taken, the probability is 0.33 (a one in three chance) that an employee drawn at random from the whole group earns between $8.00 and $8.50 an hour.

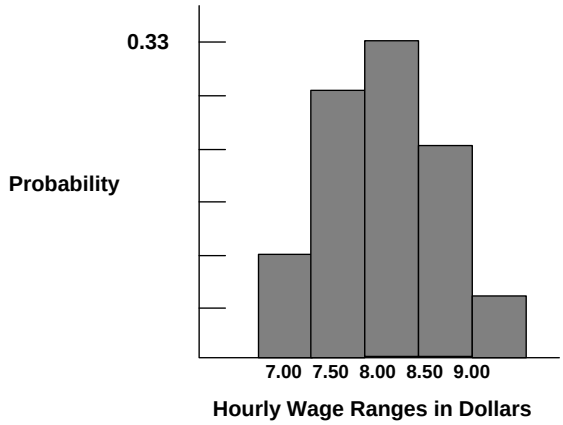


Figure 2.26 – Frequency Histogram II

Probability distributions are either discrete or continuous. Discrete probability distributions describe distinct values, usually integers, with no intermediate values and are shown as a series of vertical bars. A discrete distribution, for example, might describe the number of heads in four flips of a coin as 0, 1, 2, 3, or 4. Continuous distributions are actually mathematical abstractions because they assume the existence of every possible intermediate value between two numbers. That is, a continuous distribution assumes there is an infinite number of values between any two points in the distribution. However, in many situations, you can effectively use a continuous distribution to approximate a discrete distribution even though the continuous model does not necessarily describe the situation exactly.

Selecting the Right Probability Distribution

Plotting data is one guide to selecting a probability distribution. The following steps provide another process for selecting probability distributions that best describe the uncertain variables in your spreadsheets:

- Look at the variable in question. List everything you know about the conditions surrounding this variable. You might be able to gather valuable information about the uncertain variable from historical data. If historical data are not available, use your own judgment, based on experience, listing everything you know about the uncertain variable.
- Review the descriptions of the probability distributions.
- Select the distribution that characterizes this variable. A distribution characterizes a variable when the conditions of the distribution match those of the variable.

Monte Carlo Simulation

Monte Carlo simulation in its simplest form is a random number generator that is useful for forecasting, estimation, and risk analysis. A simulation calculates numerous scenarios of a model by repeatedly picking values from a user-predefined *probability distribution* for the uncertain variables and using those values for the model. As all those scenarios produce associated results in a model, each scenario can have a *forecast*. Forecasts are events (usually with formulas or functions) that you define as important outputs of the model. These usually are events such as totals, net profit, or gross expenses.

Simplistically, think of the Monte Carlo simulation approach as repeatedly picking golf balls out of a large basket with replacement. The size and shape of the basket depend on the distributional ***input assumption*** (e.g., a normal distribution with a mean of 100 and a standard deviation of 10, versus a uniform distribution or a triangular distribution) where some baskets

are deeper or more symmetrical than others, allowing certain balls to be pulled out more frequently than others. The number of balls pulled repeatedly depends on the number of ***trials*** simulated. For a large model with multiple related assumptions, imagine a very large basket wherein many smaller baskets reside. Each small basket has its own set of golf balls that are bouncing around. Sometimes these small baskets are linked with each other (if there is a ***correlation*** between the variables) and the golf balls are bouncing in tandem, while other times the balls are bouncing independent of one another. The balls that are picked each time from these interactions within the model (the large central basket) are tabulated and recorded, providing a ***forecast output*** result of the simulation.

With Monte Carlo simulation, Risk Simulator generates random values for each assumption's probability distribution that are totally independent. In other words, the random value selected for one trial has no effect on the next random value generated. Use Monte Carlo sampling when you want to simulate real-world what-if scenarios for your spreadsheet model.

The two following sections provide a detailed listing of the different types of discrete and continuous probability distributions that can be used in Monte Carlo simulation.

2.4 Discrete Distributions

Bernoulli or Yes/No Distribution

The Bernoulli distribution is a discrete distribution with two outcomes (e.g., head or tails, success or failure, 0 or 1). It is the binomial distribution with one trial and can be used to simulate Yes/No or Success/Failure conditions. This distribution is the fundamental building block of other more complex distributions. For instance:

- Binomial distribution: a Bernoulli distribution with higher number of *n* total trials that computes the probability of *x* successes within this total number of trials.
- Geometric distribution: a Bernoulli distribution with higher number of trials that computes the number of failures required before the first success occurs.
- Negative binomial distribution: a Bernoulli distribution with higher number of trials that computes the number of failures before the *X*th success occurs.

The mathematical constructs for the Bernoulli distribution are as follows:

$$P(n) = \begin{cases} 1-p & \text{for } x = 0 \\ p & \text{for } x = 1 \end{cases}$$

or

$$P(n) = p^x (1-p)^{1-x}$$

$$Mean = p$$

$$Standard\ Deviation = \sqrt{p(1-p)}$$

$$Skewness = \frac{1-2p}{\sqrt{p(1-p)}}$$

$$Excess\ Kurtosis = \frac{6p^2 - 6p + 1}{p(1-p)}$$

Probability of success (*p*) is the only distributional parameter. Also, it is important to note that there is only one trial in the Bernoulli distribution, and the resulting simulated value is either 0 or 1.

Input requirements:

Probability of success > 0 and < 1 (i.e., $0.0001 \leq p \leq 0.9999$).

Binomial Distribution

The binomial distribution describes the number of times a particular event occurs in a fixed number of trials, such as the number of heads in 10 flips of a coin or the number of defective items out of 50 items chosen.

Conditions

The three conditions underlying the binomial distribution are:

- For each trial, only two outcomes are possible that are mutually exclusive.
- The trials are independent—what happens in the first trial does not affect the next trial.

- The probability of an event occurring remains the same from trial to trial.

The mathematical constructs for the binomial distribution are as follows:

$$P(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{(n-x)} \quad \text{for } n > 0; x = 0, 1, 2, \ldots n; \text{and } 0 < p < 1$$

$$Mean = np$$

$$Standard\ Deviation = \sqrt{np(1-p)}$$

$$Skewness = \frac{1-2p}{\sqrt{np(1-p)}}$$

$$Excess\ Kurtosis = \frac{6p^2 - 6p + 1}{np(1-p)}$$

Probability of success (p) and the integer number of total trials (n) are the distributional parameters. The number of successful trials is denoted x. It is important to note that probability of success (p) of 0 or 1 are trivial conditions that do not require any simulations and, hence, are not allowed in the software.

Input requirements:

Probability of success > 0 and < 1 (i.e., $0.0001 \leq p \leq 0.9999$).

Number of trials ≥ 1 or positive integers and ≤ 1000 (for larger trials, use the normal distribution with the relevant computed binomial mean and standard deviation as the normal distribution's parameters).

Discrete Uniform

The discrete uniform distribution is also known as the *equally likely outcomes* distribution, where the distribution has a set of N elements and each element has the same probability. This distribution is related to the uniform distribution but its elements are discrete and not continuous.

The mathematical constructs for the discrete uniform distribution are as follows:

$$P(x) = \frac{1}{N}$$

$$Mean = \frac{N+1}{2} \quad \textit{ranked value}$$

$$Standard\ Deviation = \sqrt{\frac{(N-1)(N+1)}{12}} \quad \textit{ranked value}$$

Skewness = 0 (i.e., the distribution is perfectly symmetrical)

$$Excess\ Kurtosis = \frac{-6(N^2+1)}{5(N-1)(N+1)} \quad \textit{ranked value}$$

Input requirements:

Minimum < maximum and both must be integers (negative integers and zero are allowed).

Geometric Distribution

The geometric distribution describes the number of trials until the first successful occurrence, such as the number of times you need to spin a roulette wheel before you win.

Conditions

The three conditions underlying the geometric distribution are:

- The number of trials is not fixed.
- The trials continue until the first success.
- The probability of success is the same from trial to trial.

The mathematical constructs for the geometric distribution are as follows:

$$P(x) = p(1-p)^{x-1} \quad \text{for } 0 < p < 1 \text{ and } x = 1, 2, \ldots, n$$

$$Mean = \frac{1}{p} - 1$$

$$Standard\ Deviation = \sqrt{\frac{1-p}{p^2}}$$

$$Skewness = \frac{2-p}{\sqrt{1-p}}$$

$$Excess\ Kurtosis = \frac{p^2 - 6p + 6}{1-p}$$

Probability of success (p) is the only distributional parameter. The number of successful trials simulated is denoted x, which can only take on positive integers.

Input requirements:

Probability of success > 0 and < 1 (i.e., $0.0001 \leq p \leq 0.9999$). It is important to note that probability of success (p) of 0 or 1 are trivial conditions that do not require any simulations and, hence, are not allowed in the software.

Hypergeometric Distribution

The hypergeometric distribution is similar to the binomial distribution in that both describe the number of times a particular event occurs in a fixed number of trials. The difference is that binomial distribution trials are independent, whereas hypergeometric distribution trials change the probability for each subsequent trial and are called "trials without replacement." For example, suppose a box of manufactured parts is known to contain some defective parts. You choose a part from the box, find it is defective, and remove the part from the box. If you choose another part from the box, the probability that it is defective is somewhat lower than for the first part because you have already removed a defective part. If you had replaced the defective part, the probabilities would have remained the same, and the process would have satisfied the conditions for a binomial distribution.

Conditions

The three conditions underlying the hypergeometric distribution are:

- The total number of items or elements (the population size) is a fixed number, a finite population. The population size must be less than or equal to 1,750.
- The sample size (the number of trials) represents a portion of the population.
- The known initial probability of success in the population changes after each trial.

The mathematical constructs for the hypergeometric distribution are as follows:

$$P(x) = \frac{\dfrac{(N_x)!}{x!(N_x - x)!}\dfrac{(N - N_x)!}{(n-x)!(N - N_x - n + x)!}}{\dfrac{N!}{n!(N-n)!}} \quad \text{for } x = Max(n - (N - N_x), 0), \ldots, Min(n, N_x)$$

$$Mean = \frac{N_x n}{N}$$

$$Standard\ Deviation = \sqrt{\frac{(N - N_x)N_x n(N-n)}{N^2(N-1)}}$$

$$Skewness = \sqrt{\frac{N-1}{(N - N_x)N_x n(N-n)}}$$

Excess Kurtosis = complex function

The number of items in the population or Population Size (N), trials sampled or Sample Size (n), and number of items in the population that have the successful trait or Population Successes (N_x) are the distributional parameters. The number of successful trials is denoted x.

Input requirements:

Population Size ≥ 2 and integer.

Sample Size > 0 and integer.

Population Successes > 0 and integer.

Population Size > Population Successes.

Sample Size < Population Successes.

Population Size < 1750.

Negative Binomial Distribution

The negative binomial distribution is useful for modeling the distribution of the number of additional trials required in addition to the number of successful occurrences required (R). For instance, in order to close a total of 10 sales opportunities, how many extra sales calls would you need to make above 10 calls given some probability of success in each call? The x-axis shows the number of additional calls required or the number of failed calls. The number of trials is not fixed, the trials continue until the Rth success, and the probability of success is the same from trial to trial. Probability of success (p) and number of successes required (R) are the distributional parameters. It is essentially a *superdistribution* of the geometric and binomial

distributions. This distribution shows the probabilities of each number of trials in excess of *R* to produce the required success *R*.

Conditions

The three conditions underlying the negative binomial distribution are:

- The number of trials is not fixed.
- The trials continue until the *r*th success.
- The probability of success is the same from trial to trial.

The mathematical constructs for the negative binomial distribution are as follows:

$$P(x) = \frac{(x+r-1)!}{(r-1)!x!} p^r (1-p)^x \quad \text{for } x = r, \text{r}+1, \ldots; \text{and } 0 < p < 1$$

$$Mean = \frac{r(1-p)}{p}$$

$$Standard\ Deviation = \sqrt{\frac{r(1-p)}{p^2}}$$

$$Skewness = \frac{2-p}{\sqrt{r(1-p)}}$$

$$Excess\ Kurtosis = \frac{p^2 - 6p + 6}{r(1-p)}$$

Probability of success (*p*) and required successes (*R*) are the distributional parameters.

Input requirements:

Successes required must be positive integers > 0 and < 8000.

Probability of success > 0 and < 1 (that is, $0.0001 \leq p \leq 0.9999$). It is important to note that probability of success (*p*) of 0 or 1 are trivial conditions that do not require any simulations and, hence, are not allowed in the software.

Pascal Distribution

The Pascal distribution is useful for modeling the distribution of the number of total trials required to obtain the number of successful occurrences required. For instance, to close a total of 10 sales opportunities, how many total sales calls would you need to make given some probability of success in each call? The *x*-axis shows the total number of calls required, which includes successful and failed calls. The number of trials is not fixed, the trials continue until the *R*th success, and the probability of success is the same from trial to trial. Pascal distribution is related to the negative binomial distribution. Negative binomial distribution computes the number of events required in addition to the number of successes required given some probability (in other words, the total failures), whereas the Pascal distribution computes the total number of events required (in other words, the sum of failures and successes) to achieve the successes required given some probability. Successes required and probability, are the two distributional parameters.

Conditions

The three conditions underlying the negative binomial distribution are:

- The number of trials is not fixed.
- The trials continue until the *r*th success.
- The probability of success is the same from trial to trial.

The mathematical constructs for the Pascal distribution are shown below:

$$f(x) = \begin{cases} \dfrac{(x-1)!}{(x-s)!(s-1)!} p^S (1-p)^{X-S} & \text{for all } x \geq s \\ 0 \text{ otherwise} \end{cases}$$

$$F(x) = \begin{cases} \sum_{x=1}^{k} \dfrac{(x-1)!}{(x-s)!(s-1)!} p^S (1-p)^{X-S} & \text{for all } x \ \geq \ s \\ 0 \text{ otherwise} \end{cases}$$

$$Mean = \frac{s}{p}$$

$$Standard\ Deviation\ = \sqrt{s(1-p)p^2}$$

$$Skewness\ = \frac{2-p}{\sqrt{r(1-p)}}$$

$$Excess\ Kurtosis\ = \frac{p^2 - 6p + 6}{r(1-p)}$$

Successes Required and Probability are the distributional parameters.

Input requirements:

Successes required > 0 and is an integer.

0 ≤ Probability ≤ 1.

Poisson Distribution

The Poisson distribution describes the number of times an event occurs in a given interval, such as the number of telephone calls per minute or the number of errors per page in a document.

Conditions

The three conditions underlying the Poisson distribution are:

- The number of possible occurrences in any interval is unlimited.
- The occurrences are independent. The number of occurrences in one interval does not affect the number of occurrences in other intervals.
- The average number of occurrences must remain the same from interval to interval.

The mathematical constructs for the Poisson are as follows:

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \quad \text{for } x \text{ and } \lambda > 0$$

$$Mean = \lambda$$

Standard Deviation = $\sqrt{\lambda}$

$$Skewness = \frac{1}{\sqrt{\lambda}}$$

Excess Kurtosis = $\frac{1}{\lambda}$

Rate, or Lambda (λ), is the only distributional parameter.

Input requirements:

Rate > 0 and ≤ 1000 (i.e., 0.0001 ≤ rate ≤ 1000).

2.5 Continuous Distributions

Arcsine Distribution

The arcsine distribution is U-shaped and is a special case of the beta distribution when both shape and scale are equal to 0.5. Values close to the minimum and maximum have high probabilities of occurrence whereas values between these two extremes have very small probabilities of occurrence. Minimum and maximum are the distributional parameters.

The mathematical constructs for the Arcsine distribution are shown below. The probability density function (PDF) is denoted *f(x)* and the cumulative distribution function (CDF) is denoted *F(x)*.

$$f(x) = \begin{cases} \dfrac{1}{\pi\sqrt{x(1-x)}} & for\ 0 \le x \le 1 \\ 0 & otherwise \end{cases}$$

$$F(x) = \begin{cases} 0 & x < 0 \\ \dfrac{2}{\pi}\sin^{-1}(\sqrt{x}) & for\ 0 \le x \le 1 \\ 1 & x > 1 \end{cases}$$

$$Mean = \frac{Min + Max}{2}$$

$$Standard\ Deviation = \sqrt{\frac{(Max - Min)^2}{8}}$$

Skewness = 0 for all inputs

Excess Kurtosis = 1.5 for all inputs

Minimum and maximum are the distributional parameters.

Input requirements:

Maximum > minimum (either input parameter can be positive, negative, or zero).

Beta Distribution

The beta distribution is very flexible and is commonly used to represent variability over a fixed range. One of the more important applications of the beta distribution is its use as a conjugate distribution for the parameter of a Bernoulli distribution. In this application, the beta distribution is used to represent the uncertainty in the probability of occurrence of an event. It is also used to describe empirical data and predict the random behavior of percentages and fractions, as the range of outcomes is typically between 0 and 1.

The value of the beta distribution lies in the wide variety of shapes it can assume when you vary the two parameters, alpha and beta. If the parameters are equal, the distribution is symmetrical. If either parameter is 1 and the other parameter is greater than 1, the distribution is J-shaped. If alpha is less than beta, the distribution is said to be positively skewed (most of the values are near the minimum value). If alpha is greater than beta, the distribution is negatively skewed (most of the values are near the maximum value).

The mathematical constructs for the beta distribution are as follows:

$$f(x) = \frac{(x)^{(\alpha-1)}(1-x)^{(\beta-1)}}{\left[\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\right]} \quad \text{for } \alpha > 0;\ \beta > 0;\ x > 0$$

$$Mean = \frac{\alpha}{\alpha+\beta}$$

$$Standard\ Deviation = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(1+\alpha+\beta)}}$$

$$Skewness = \frac{2(\beta-\alpha)\sqrt{1+\alpha+\beta}}{(2+\alpha+\beta)\sqrt{\alpha\beta}}$$

$$Excess\ Kurtosis = \frac{3(\alpha+\beta+1)[\alpha\beta(\alpha+\beta-6)+2(\alpha+\beta)^2]}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)} - 3$$

Alpha (α) and beta (β) are the two distributional shape parameters, and Γ is the Gamma function.

Conditions

The two conditions underlying the beta distribution are:

- The uncertain variable is a random value between 0 and a positive value.
- The shape of the distribution can be specified using two positive values.

Input requirements:

Alpha and beta both > 0 and can be any positive value.

Beta 3 and Beta 4 Distributions

The original Beta distribution only takes two inputs, Alpha and Beta shape parameters. However, the output of the simulated value is between 0 and 1. In the Beta 3 distribution, we add an extra parameter called Location or Shift, where we are not free to move away from this 0 to 1 output limitation, therefore the Beta 3 distribution is also known as a Shifted Beta distribution. Similarly, the Beta 4 distribution adds two input parameters, Location or Shift, and Factor. The original beta distribution is multiplied by the factor and shifted by the location, and, therefore the Beta 4 is also known as the Multiplicative Shifted Beta distribution.

The mathematical constructs for the Beta 3 and Beta 4 distributions are based on those in the Beta distribution, with the relevant shifts and factorial multiplication (e.g., the PDF and CDF will be adjusted by the shift and factor, and some of the moments, such as the mean, will similarly be affected; the standard deviation, in contrast, is only affected by the factorial multiplication, whereas the remaining moments are not affected at all).

Input requirements:

Location >=< 0 (location can take on any positive or negative value including zero).

Factor > 0.

Cauchy Distribution, or Lorentzian or Breit-Wigner Distribution

The Cauchy distribution, also called the Lorentzian or Breit-Wigner distribution, is a continuous distribution describing resonance behavior. It also describes the distribution of horizontal distances at which a line segment tilted at a random angle cuts the *x*-axis.

The mathematical constructs for the cauchy or Lorentzian distribution are as follows:

$$f(x) = \frac{1}{\pi} \frac{\gamma / 2}{(x-m)^2 + \gamma^2 / 4}$$

The Cauchy distribution is a special case because it does not have any theoretical moments (mean, standard deviation, skewness, and kurtosis) as they are all undefined.

Mode location (α) and scale (β) are the only two parameters in this distribution. The location parameter specifies the peak or mode of the distribution, while the scale parameter specifies the half-width at half-maximum of the distribution. In addition, the mean and variance of a Cauchy, or Lorentzian, distribution are undefined.

In addition, the Cauchy distribution is the Student's T distribution with only 1 degree of freedom. This distribution is also constructed by taking the ratio of two standard normal distributions (normal distributions with a mean of zero and a variance of one) that are independent of one another.

Input requirements: Location (Alpha) can be any value. Scale (Beta) > 0 and can be any positive value.

Chi-Square Distribution

The chi-square distribution is a probability distribution used predominantly in hypothesis testing, and is related to the gamma and standard normal distributions. For instance, the sum of independent normal distributions is distributed as a chi-square (χ^2) with *k* degrees of freedom:

$$Z_1^2 + Z_2^2 + \ldots + Z_k^2 \overset{d}{\sim} \chi_k^2$$

The mathematical constructs for the chi-square distribution are as follows:

$$f(x) = \frac{0.5^{-k/2}}{\Gamma(k/2)} x^{k/2-1} e^{-x/2} \text{ for all } x > 0$$

Mean = *k*

Standard Deviation = $\sqrt{2k}$

Skewness = $2\sqrt{\frac{2}{k}}$

Excess Kurtosis = $\frac{12}{k}$

Γ is the gamma function. Degrees of freedom, *k*, is the only distributional parameter.

The chi-square distribution can also be modeled using a gamma distribution by setting the

Shape parameter equal to $\frac{k}{2}$ and the scale equal to $2S^2$ where *S* is the scale.

Input requirements: Degrees of freedom > 1 and must be an integer < 300.

Cosine Distribution

The cosine distribution looks like a logistic distribution where the median value between the minimum and maximum have the highest peak or mode, carrying the maximum probability of occurrence, while the extreme tails close to the minimum and maximum values have lower probabilities. Minimum and maximum are the distributional parameters.

The mathematical constructs for the Cosine distribution are shown below:

$$f(x) = \begin{cases} \frac{1}{2b}\cos\left[\frac{x-a}{b}\right] & for\ \min \le x \le \max \\ 0 & otherwise \end{cases}$$

$$where\ a = \frac{\min + \max}{2}\ and\ b = \frac{\max - \min}{\pi}$$

$$F(x) = \begin{cases} \frac{1}{2}\left[1 + \sin\left(\frac{x-a}{b}\right)\right] & for\ \min \le x \le \max \\ 1 & for\ x > max \end{cases}$$

$$Mean = \frac{Min + Max}{2}$$

$$Standard\ Deviation = \sqrt{\frac{(Max - Min)^2(\pi^2 - 8)}{4\pi^2}}$$

Skewness is always equal to 0

$$Excess\ Kurtosis = \frac{6(90 - \pi^4)}{5(\pi^2 - 6)^2}$$

Minimum and maximum are the distributional parameters.

Input requirements:

Maximum > minimum (either input parameter can be positive, negative, or zero).

Double Log Distribution

The double log distribution looks like the Cauchy distribution where the central tendency is peaked and carries the maximum value probability density but declines faster the further it gets away from the center, creating a symmetrical distribution with an extreme peak in between the minimum and maximum values. Minimum and maximum are the distributional parameters.

The mathematical constructs for the Double Log distribution are shown below:

$$f(x) = \begin{cases} \frac{-1}{2b}\ln\left(\frac{|x-a|}{b}\right) & for\ \min \le x \le \max \\ 0 & otherwise \end{cases}$$

$$where\ a = \frac{\min + \max}{2}\ and\ b = \frac{\max - \min}{2}$$

$$F(x) = \begin{cases} \frac{1}{2} - \left(\frac{|x-a|}{2b}\right)\left[1 - \ln\left(\frac{|x-a|}{b}\right)\right] & for\ \min \le x \le a \\ \frac{1}{2} + \left(\frac{|x-a|}{2b}\right)\left[1 - \ln\left(\frac{|x-a|}{b}\right)\right] & for\ a \le x \le \max \end{cases}$$

$$Mean = \frac{Min + Max}{2}$$

$$Standard\ Deviation = \sqrt{\frac{(Max - Min)^2}{36}}$$

Skewness is always equal to 0

Excess Kurtosis is a complex function and not easily represented

Minimum and maximum are the distributional parameters.

Input requirements:

Maximum > minimum (either input parameter can be positive, negative, or zero).

Erlang Distribution

The Erlang distribution is the same as the Gamma distribution with the requirement that the Alpha or shape parameter must be a positive integer. An example application of the Erlang distribution is the calibration of the rate of transition of elements through a system of compartments. Such systems are widely used in biology and ecology (e.g., in epidemiology, an individual may progress at an exponential rate from being healthy to becoming a disease carrier, and continue exponentially from being a carrier to being infectious). Alpha (also known as shape) and Beta (also known as scale) are the distributional parameters.

The mathematical constructs for the Erlang distribution are shown below:

$$f(x) = \begin{cases} \dfrac{\left(\dfrac{x}{\beta}\right)^{\alpha-1} e^{-x/\beta}}{\beta(\alpha-1)} & for\ x \geq 0 \\ 0 & otherwise \end{cases}$$

$$F(x) = \begin{cases} 1 - e^{-x/\beta} \sum_{i=0}^{\alpha-1} \dfrac{(x/\beta)^i}{i!} & for\ x \geq 0 \\ 0 & otherwise \end{cases}$$

$$Mean = \alpha\beta$$

$$Standard\ Deviation = \sqrt{\alpha\beta^2}$$

$$Skew = \frac{2}{\sqrt{\alpha}}$$

$$Excess\ Kurtosis = \frac{6}{\alpha} - 3$$

Alpha and Beta are the distributional parameters.

Input requirements:

Alpha (Shape) > 0 and is an Integer

Beta (Scale) > 0

Exponential Distribution

The exponential distribution is widely used to describe events recurring at random points in time, such as the time between failures of electronic equipment or the time between arrivals at a service booth. It is related to the Poisson distribution, which describes the number of occurrences of an event in a given interval of time. An important characteristic of the exponential distribution is the "memoryless" property, which means that the future lifetime of a given object has the same distribution regardless of the time it existed. In other words, time has no effect on future outcomes.

Conditions

The condition underlying the exponential distribution is:

- The exponential distribution describes the amount of time between occurrences.

The mathematical constructs for the exponential distribution are as follows:

$$f(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0; \lambda > 0$$

$$Mean = \frac{1}{\lambda}$$

$$\text{Standard Deviation} = \frac{1}{\lambda}$$

Skewness = 2 (this value applies to all success rate λ inputs)

Excess Kurtosis = 6 (this value applies to all success rate λ inputs)

Success rate (λ) is the only distributional parameter. The number of successful trials is denoted x.

Input requirements: Rate > 0.

Exponential 2 Distribution

The Exponential 2 distribution uses the same constructs as the original Exponential distribution but adds a Location or Shift parameter. The Exponential distribution starts from a minimum value of 0, whereas this Exponential 2 or Shifted Exponential, distribution shifts the starting location to any other value.

Rate, or Lambda, and Location, or Shift, are the distributional parameters.

Input requirements:

Rate (Lambda) > 0.

Location can be any positive or negative value including zero.

Extreme Value Distribution, or Gumbel Distribution

The extreme value distribution (Type 1) is commonly used to describe the largest value of a response over a period of time, for example, in flood flows, rainfall, and earthquakes. Other applications include the breaking strengths of materials, construction design, and aircraft loads and tolerances. The extreme value distribution is also known as the Gumbel distribution.

The mathematical constructs for the extreme value distribution are as follows:

$$f(x) = \frac{1}{\beta} z e^{-Z} \text{ where } z = e^{\frac{x-\alpha}{\beta}} \quad \text{for } \beta > 0; \text{and any value of } x \text{ and } \alpha$$

Mean = $\alpha + 0.577215\beta$

Standard Deviation = $\sqrt{\frac{1}{6}\pi^2\beta^2}$

Skewness = $\frac{12\sqrt{6}(1.2020569)}{\pi^3} = 1.13955$ (this applies for all values of mode and scale)

Excess Kurtosis = 5.4 (this applies for all values of mode and scale)

Mode (α) and scale (β) are the distributional parameters.

Calculating Parameters

There are two standard parameters for the extreme value distribution: mode and scale. The mode parameter is the most likely value for the variable (the highest point on the probability distribution). After you select the mode parameter, you can estimate the scale parameter. The scale parameter is a number greater than 0. The larger the scale parameter, the greater the variance.

Input requirements:

Mode Alpha can be any value.

Scale Beta > 0.

F Distribution, or Fisher-Snedecor Distribution

The F distribution, also known as the Fisher-Snedecor distribution, is another continuous distribution used most frequently for hypothesis testing. Specifically, it is used to test the statistical difference between two variances in analysis of variance tests and likelihood ratio tests. The F distribution with the numerator degree of freedom *n* and denominator degree of freedom *m* is related to the chi-square distribution in that:

$$\frac{\chi_n^2 / n}{\chi_m^2 / m} \overset{d}{\sim} F_{n,m}$$

Mean = $\frac{m}{m-2}$

Standard Deviation = $\frac{2m^2(m+n-2)}{n(m-2)^2(m-4)}$ *for all m > 4*

Skewness = $\frac{2(m+2n-2)}{m-6}\sqrt{\frac{2(m-4)}{n(m+n-2)}}$

Excess Kurtosis = $\frac{12(-16+20m-8m^2+m^3+44n-32mn+5m^2n-22n^2+5mn^2}{n(m-6)(m-8)(n+m-2)}$

The numerator degree of freedom *n* and denominator degree of freedom *m* are the only distributional parameters.

Input requirements:

Degrees of freedom numerator & degrees of freedom denominator must both be integers > 0

Gamma Distribution (Erlang Distribution)

The gamma distribution applies to a wide range of physical quantities and is related to other distributions: lognormal, exponential, Pascal, Erlang, Poisson, and chi-square. It is used in meteorological processes to represent pollutant concentrations and precipitation quantities. The gamma distribution is also used to measure the time between the occurrence of events when the event process is not completely random. Other applications of the gamma distribution include inventory control, economic theory, and insurance risk theory.

Conditions

The gamma distribution is most often used as the distribution of the amount of time until the rth occurrence of an event in a Poisson process. When used in this fashion, the three conditions underlying the gamma distribution are:

- The number of possible occurrences in any unit of measurement is not limited to a fixed number.
- The occurrences are independent. The number of occurrences in one unit of measurement does not affect the number of occurrences in other units.
- The average number of occurrences must remain the same from unit to unit.

The mathematical constructs for the gamma distribution are as follows:

$$f(x) = \frac{\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta} \quad \text{with any value of } \alpha > 0 \text{ and } \beta > 0$$

$$Mean = \alpha\beta$$

$$Standard\ Deviation = \sqrt{\alpha\beta^2}$$

$$Skewness = \frac{2}{\sqrt{\alpha}}$$

$$Excess\ Kurtosis = \frac{6}{\alpha}$$

Shape parameter alpha (α) and scale parameter beta (β) are the distributional parameters, and Γ is the Gamma function. When the alpha parameter is a positive integer, the gamma distribution is called the Erlang distribution, used to predict waiting times in queuing systems, where the Erlang distribution is the sum of independent and identically distributed random variables each having a memoryless exponential distribution. Setting n as the number of these random variables, the mathematical construct of the Erlang distribution is:

$$f(x) = \frac{x^{n-1}e^{-x}}{(n-1)!} \quad \text{for all } x > 0 \text{ and all positive integers of } n$$

Input requirements:

Scale beta > 0 and can be any positive value.

Shape alpha ≥ 0.05 and any positive value.

Location can be any value.

Laplace Distribution

The Laplace distribution is also sometimes called the double exponential distribution because it can be constructed with two exponential distributions (with an additional location parameter) spliced together back-to-back, creating an unusual peak in the middle. The probability density function of the Laplace distribution is reminiscent of the normal distribution. However, whereas the normal distribution is expressed in terms of the squared difference from the mean, the Laplace density is expressed in terms of the absolute difference from the mean, making the Laplace distribution's tails fatter than those of the normal distribution. When the location parameter is set to zero, the Laplace distribution's random variable is exponentially distributed with an inverse of the scale parameter. Alpha (also known as location) and Beta (also known as scale) are the distributional parameters.

The mathematical constructs for the Laplace distribution are shown below:

$$f(x) = \frac{1}{2\beta}\exp\left(-\frac{|x-\alpha|}{\beta}\right)$$

$$F(x) = \begin{cases} \frac{1}{2}\exp\left[\frac{x-\alpha}{\beta}\right] & \text{when } x < \alpha \\ 1-\frac{1}{2}\exp\left[-\frac{x-\alpha}{\beta}\right] & \text{when } x \geq \alpha \end{cases}$$

$$Mean = \alpha$$

$$Standard\ Deviation\ =\ 1.4142\beta$$

Skewness is always equal to 0 as it is a symmetrical distribution

Excess Kurtosis is always equal to 3

Input requirements:

Alpha (Location) can take on any positive or negative value including zero.

Beta (Scale) > 0.

Logistic Distribution

The logistic distribution is commonly used to describe growth, that is, the size of a population expressed as a function of a time variable. It also can be used to describe chemical reactions and the course of growth for a population or individual.

The mathematical constructs for the logistic distribution are as follows:

$$f(x) = \frac{e^{\frac{\alpha-x}{\beta}}}{\beta\left[1+e^{\frac{\alpha-x}{\beta}}\right]^2} \quad \text{for any value of } \alpha \text{ and } \mu$$

$$Mean\ = \alpha$$

$$Standard\ Deviation\ = \sqrt{\frac{1}{3}\pi^2\beta^2}$$

Skewness = 0 (this applies to all mean and scale inputs)

Excess Kurtosis = 1.2 (this applies to all mean and scale inputs)

Mean (α) and scale (β) are the distributional parameters.

Calculating Parameters

There are two standard parameters for the logistic distribution: mean and scale. The mean parameter is the average value, which for this distribution is the same as the mode because this is a symmetrical distribution. After you select the mean parameter, you can estimate the scale parameter. The scale parameter is a number greater than 0. The larger the scale parameter, the greater the variance.

Input requirements:

Scale Beta > 0 and can be any positive value.

Mean Alpha can be any value.

Lognormal Distribution

The lognormal distribution is widely used in situations where values are positively skewed, for example, in financial analysis for security valuation or in real estate for property valuation, and where values cannot fall below zero.

Stock prices are usually positively skewed rather than normally (symmetrically) distributed. Stock prices exhibit this trend because they cannot fall below the lower limit of zero but might increase to any price without limit. Similarly, real estate prices illustrate positive skewness as property values cannot become negative.

Conditions

The three conditions underlying the lognormal distribution are:

- The uncertain variable can increase without limits but cannot fall below zero.
- The uncertain variable is positively skewed, with most of the values near the lower limit.
- The natural logarithm of the uncertain variable yields a normal distribution.

Generally, if the coefficient of variability is greater than 30%, use a lognormal distribution. Otherwise, use the normal distribution.

The mathematical constructs for the lognormal distribution are as follows:

$$f(x) = \frac{1}{x\sqrt{2\pi}\ln(\sigma)} e^{-\frac{[\ln(x)-\ln(\mu)]^2}{2[\ln(\sigma)]^2}} \quad \text{for } x > 0;\ \mu > 0 \text{ and } \sigma > 0$$

$$Mean = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

$$Standard\ Deviation = \sqrt{\exp(\sigma^2 + 2\mu)\left[\exp(\sigma^2) - 1\right]}$$

$$Skewness = \left[\sqrt{\exp(\sigma^2) - 1}\right](2 + \exp(\sigma^2))$$

$$Excess\ Kurtosis = \exp(4\sigma^2) + 2\exp(3\sigma^2) + 3\exp(2\sigma^2) - 6$$

Mean (μ) and standard deviation (σ) are the distributional parameters.

Input requirements:

Mean and standard deviation both > 0 and can be any positive value.

Lognormal Parameter Sets

By default, the lognormal distribution uses the arithmetic mean and standard deviation. For applications for which historical data are available, it is more appropriate to use either the logarithmic mean and standard deviation, or the geometric mean and standard deviation.

Lognormal 3 Distribution

The Lognormal 3 distribution uses the same constructs as the original Lognormal distribution but adds a Location, or Shift, parameter. The Lognormal distribution starts from a minimum value of 0, whereas this Lognormal 3, or Shifted Lognormal distribution shifts the starting location to any other value.

Mean, Standard Deviation, and Location (Shift) are the distributional parameters.

Input requirements:

Mean > 0.

Standard Deviation > 0.

Location can be any positive or negative value including zero.

Normal Distribution

The normal distribution is the most important distribution in probability theory because it describes many natural phenomena, such as people's IQs or heights. Decision makers can use the normal distribution to describe uncertain variables such as the inflation rate or the future price of gasoline.

Conditions

The three conditions underlying the normal distribution are:

- Some value of the uncertain variable is the most likely (the mean of the distribution).
- The uncertain variable could as likely be above the mean as it could be below the mean (symmetrical about the mean).
- The uncertain variable is more likely to be in the vicinity of the mean than further away.

The mathematical constructs for the normal distribution are as follows:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for all values of } x$$

and μ; while σ > 0

$$Mean = \mu$$

$$Standard\ Deviation = \sigma$$

Skewness = 0 (this applies to all inputs of mean and standard deviation)

Excess Kurtosis = 0 (this applies to all inputs of mean and standard deviation)

Mean (μ) and standard deviation (σ) are the distributional parameters.

Input requirements:

Standard deviation > 0 and can be any positive value.

Mean can take on any value.

Parabolic Distribution

The parabolic distribution is a special case of the beta distribution when Shape = Scale = 2. Values close to the minimum and maximum have low probabilities of occurrence, whereas values between these two extremes have higher probabilities or occurrence. Minimum and maximum are the distributional parameters.

The mathematical constructs for the Parabolic distribution are shown below:

$$f(x) = \frac{(x)^{(\alpha-1)}(1-x)^{(\beta-1)}}{\left[\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\right]} \quad \text{for } \alpha > 0;\ \beta > 0;\ x > 0$$

Where the functional form above is for a Beta distribution, and for a Parabolic function, we set Alpha = Beta = 2 and a shift of location in Minimum, with a multiplicative factor of (Maximum – Minimum).

$$Mean = \frac{Min + Max}{2}$$

$$Standard\ Deviation = \sqrt{\frac{(Max - Min)^2}{20}}$$

$$Skewness = 0$$

$$Excess\ Kurtosis = -0.8571$$

Minimum and Maximum are the distributional parameters.

Input requirements:

Maximum > minimum (either input parameter can be positive, negative, or zero).

Pareto Distribution

The Pareto distribution is widely used for the investigation of distributions associated with such empirical phenomena as city population sizes, the occurrence of natural resources, the size of companies, personal incomes, stock price fluctuations, and error clustering in communication circuits.

The mathematical constructs for the Pareto are as follows:

$$f(x) = \frac{\beta L^{\beta}}{x^{(1+\beta)}} \quad \text{for } x > L$$

$$\text{mean} = \frac{\beta L}{\beta - 1}$$

$$\text{standard deviation} = \sqrt{\frac{\beta L^2}{(\beta-1)^2(\beta-2)}}$$

$$skewness = \sqrt{\frac{\beta-2}{\beta}}\left[\frac{2(\beta+1)}{\beta-3}\right]$$

$$excess\ kurtosis = \frac{6(\beta^3 + \beta^2 - 6\beta - 2)}{\beta(\beta-3)(\beta-4)}$$

Shape (α) and Location (β) are the distributional parameters.

Calculating Parameters

There are two standard parameters for the Pareto distribution: location and shape. The location parameter is the lower bound for the variable. After you select the location parameter, you can estimate the shape parameter. The shape parameter is a number greater than 0, usually greater than 1. The larger the shape parameter, the smaller the variance and the thicker the right tail of the distribution.

Input requirements:

Location > 0 and can be any positive value

Shape ≥ 0.05.

Pearson V Distribution

The Pearson V distribution is related to the Inverse Gamma distribution, where it is the reciprocal of the variable distributed according to the Gamma distribution. Pearson V distribution is also used to model time delays where there is almost certainty of some minimum delay and the maximum delay is unbounded, for example, delay in arrival of emergency services and time to repair a machine. Alpha (also known as shape) and Beta (also known as scale) are the distributional parameters.

The mathematical constructs for the Pearson V distribution are shown below:

$$f(x) = \frac{x^{-(\alpha+1)} e^{-\beta/x}}{\beta^{-\alpha}\Gamma(\alpha)}$$

$$F(x) = \frac{\Gamma(\alpha, \beta / x)}{\Gamma(\alpha)}$$

$$Mean = \frac{\beta}{\alpha - 1}$$

$$Standard\ Deviation = \sqrt{\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}}$$

$$Skew = \frac{4\sqrt{\alpha-2}}{\alpha-3}$$

$$Excess\ Kurtosis = \frac{30\alpha - 66}{(\alpha-3)(\alpha-4)} - 3$$

Input requirements:

Alpha (Shape) > 0.

Beta (Scale) > 0.

Pearson VI Distribution

The Pearson VI distribution is related to the Gamma distribution, where it is the rational function of two variables distributed according to two Gamma distributions. Alpha 1 (also known as shape 1), Alpha 2 (also known as shape 2), and Beta (also known as scale) are the distributional parameters.

The mathematical constructs for the Pearson VI distribution are shown below:

$$f(x) = \frac{(x/\beta)^{\alpha_1 - 1}}{\beta\ \mathrm{B}(\alpha_1, \alpha_2)[1 + (x/\beta)]^{\alpha_1 + \alpha_2}}$$

$$F(x) = F_B\left(\frac{x}{x+\beta}\right)$$

$$Mean = \frac{\beta\alpha_1}{\alpha_2 - 1}$$

$$Standard\ Deviation = \sqrt{\frac{\beta^2\alpha_1(\alpha_1 + \alpha_2 - 1)}{(\alpha_2 - 1)^2(\alpha_2 - 2)}}$$

$$Skew = 2\sqrt{\frac{\alpha_2 - 2}{\alpha_1(\alpha_1 + \alpha_2 - 1)}}\left[\frac{2\alpha_1 + \alpha_2 - 1}{\alpha_2 - 3}\right]$$

$$Excess\ Kurtosis = \frac{3(\alpha_2 - 2)}{(\alpha_2 - 3)(\alpha_2 - 4)}\left[\frac{2(\alpha_2 - 1)^2}{\alpha_1(\alpha_1 + \alpha_2 - 1)} + (\alpha_2 + 5)\right] - 3$$

Input requirements:

Alpha 1 (Shape 1) > 0.

Alpha 2 (Shape 2) > 0.

Beta (Scale) > 0.

PERT Distribution

The PERT distribution is widely used in project and program management to define the worst-case, nominal-case, and best-case scenarios of project completion time. It is related to the Beta and Triangular distributions. PERT distribution can be used to identify risks in project and cost models based on the likelihood of meeting targets and goals across any number of project components using minimum, most likely, and maximum values, but it is designed to generate a distribution that more closely resembles realistic probability distributions. The PERT distribution can provide a close fit to the normal or lognormal distributions. Like the triangular distribution, the PERT distribution emphasizes the "most likely" value over the minimum and maximum estimates. However, unlike the triangular distribution, the PERT distribution constructs a smooth curve that places progressively more emphasis on values around (near) the most likely value, in favor of values around the edges. In practice, this means that we "trust" the estimate for the most likely value, and we believe that even if it is not exactly accurate (as estimates seldom are), we have an expectation that the resulting value will be close to that estimate. Assuming that many real-world phenomena are normally distributed, the appeal of the PERT distribution is that it produces a curve similar to the normal curve in shape, without knowing the precise parameters of the related normal curve. Minimum, Most Likely, and Maximum are the distributional parameters.

The mathematical constructs for the PERT distribution are shown below:

$$f(x) = \frac{(x-\min)^{A1-1}(\max - x)^{A2-1}}{B(A1, A2)(\max - \min)^{A1+A2-1}}$$

$$where\ A1 = 6\left[\frac{\frac{\min + 4(\text{likely}) + \max}{6} - \min}{\max - \min}\right] and\ A2 = 6\left[\frac{\max - \frac{\min + 4(\text{likely}) + \max}{6}}{\max - \min}\right]$$

and B is the Beta function

$$Mean = \frac{Min + 4Mode + Max}{6}$$

$$Standard\ Deviation = \sqrt{\frac{(\mu - Min)(Max - \mu)}{7}}$$

$$Skew = \sqrt{\frac{7}{(\mu - Min)(Max - \mu)}}\left(\frac{Min + Max - 2\mu}{4}\right)$$

Input requirements:

Minimum ≤ Most Likely ≤ Maximum and can be positive, negative, or zero.

Power Distribution

The Power distribution is related to the exponential distribution in that the probability of small outcomes is large but exponentially decreases as the outcome value increases. Alpha (also known as shape) is the only distributional parameter.

The mathematical constructs for the Power distribution are shown below:

$$f(x) = \alpha x^{\alpha - 1}$$

$$F(x) = x^{\alpha}$$

$$Mean = \frac{\alpha}{1+\alpha}$$

$$Standard\ Deviation = \sqrt{\frac{\alpha}{(1+\alpha)^2(2+\alpha)}}$$

$$Skew = \sqrt{\frac{\alpha+2}{\alpha}}\left(\frac{2(\alpha-1)}{\alpha+3}\right)$$

Excess Kurtosis is a complex function and cannot be readily computed

Input requirements: Alpha > 0.

Power 3 Distribution

The Power 3 distribution uses the same constructs as the original Power distribution but adds a Location, or Shift, parameter, and a multiplicative Factor parameter. The Power distribution starts from a minimum value of 0, whereas this Power 3, or Shifted Multiplicative Power, distribution shifts the starting location to any other value.

Alpha, Location or Shift, and Factor are the distributional parameters.

Input requirements:

Alpha > 0.05.

Location, or Shift, can be any positive or negative value including zero.

Factor > 0.

Student's t Distribution

The Student's t distribution is the most widely used distribution in hypothesis test. This distribution is used to estimate the mean of a normally distributed population when the sample size is small to test the statistical significance of the difference between two sample means or confidence intervals for small sample sizes.

The mathematical constructs for the t distribution are as follows:

$$f(t) = \frac{\Gamma[(r+1)/2]}{\sqrt{r\pi}\,\Gamma[r/2]}(1+t^2/r)^{-(r+1)/2}$$

Mean = 0 (this applies to all degrees of freedom *r* except if the distribution is shifted to another nonzero central location)

Standard Deviation = $\sqrt{\frac{r}{r-2}}$

Skewness = 0 (this applies to all degrees of freedom r)

Excess Kurtosis = $\frac{6}{r-4}$ for all $r > 4$

where $t = \frac{x-\bar{x}}{s}$ and Γ is the gamma function.

Degrees of freedom *r* is the only distributional parameter.

The t distribution is related to the F distribution as follows: the square of a value of *t* with *r* degrees of freedom is distributed as F with 1 and *r* degrees of freedom. The overall shape of the probability density function of the t distribution also resembles the bell shape of a normally distributed variable with mean 0 and variance 1, except that it is a bit lower and wider or is leptokurtic (fat tails at the ends and peaked center). As the number of degrees of freedom grows (say, above 30), the t distribution approaches the normal distribution with mean 0 and variance 1.

Input requirements:

Degrees of freedom ≥ 1 and must be an integer.

Triangular Distribution

The triangular distribution describes a situation where you know the minimum, maximum, and most likely values to occur. For example, you could describe the number of cars sold per week when past sales show the minimum, maximum, and usual number of cars sold.

Conditions

The three conditions underlying the triangular distribution are:

- The minimum number of items is fixed.
- The maximum number of items is fixed.
- The most likely number of items falls between the minimum and maximum values, forming a triangular-shaped distribution, which shows that values near the minimum and maximum are less likely to occur than those near the most-likely value.

The mathematical constructs for the triangular distribution are as follows:

$$f(x) = \begin{cases} \dfrac{2(x - Min)}{(Max - Min)(Likely - \min)} & \text{for } Min < x < Likely \\ \dfrac{2(Max - x)}{(Max - Min)(Max - Likely)} & \text{for } Likely < x < \text{Max} \end{cases}$$

$$Mean = \frac{1}{3}(Min + Likely + Max)$$

$$Standard\ Deviation = \sqrt{\frac{1}{18}(Min^2 + Likely^2 + Max^2 - Min\,Max - Min\,Likely - Max\,Likely)}$$

$$Skewness = \frac{\sqrt{2}(Min + Max - 2Likely)(2Min - Max - Likely)(Min - 2Max + Likely)}{5(Min^2 + Max^2 + Likely^2 - MinMax - MinLikely - MaxLikely)^{3/2}}$$

Excess Kurtosis = –0.6 (this applies to all inputs of Min, Max, and Likely)

Minimum value (*Min*), most-likely value (*Likely*), and maximum value (*Max*) are the distributional parameters.

Input requirements:

Min ≤ Most Likely ≤ Max and can take any value.

However, Min < Max and can take any value.

Uniform Distribution

With the uniform distribution, all values fall between the minimum and maximum and occur with equal likelihood.

Conditions

The three conditions underlying the uniform distribution are:

- The minimum value is fixed.
- The maximum value is fixed.
- All values between the minimum and maximum occur with equal likelihood.

The mathematical constructs for the uniform distribution are as follows:

$$f(x) = \frac{1}{Max - Min} \quad \text{for all values such that } Min < Max$$

$$Mean = \frac{Min + Max}{2}$$

$$Standard\ Deviation = \sqrt{\frac{(Max - Min)^2}{12}}$$

Skewness = 0 (this applies to all inputs of *Min* and *Max*)

Excess Kurtosis = –1.2 (this applies to all inputs of Min and Max)

Maximum value (*Max*) and minimum value (*Min*) are the distributional parameters.

Input requirements:

Min < Max and can take any value.

Weibull Distribution (Rayleigh Distribution)

The Weibull distribution describes data resulting from life and fatigue tests. It is commonly used to describe failure time in reliability studies as well as the breaking strengths of materials in reliability and quality control tests. Weibull distributions are also used to represent various physical quantities, such as wind speed.

The Weibull distribution is a family of distributions that can assume the properties of several other distributions. For example, depending on the shape parameter you define, the Weibull distribution can be used to model the exponential and Rayleigh distributions, among others. The Weibull distribution is very flexible. When the Weibull shape parameter is equal to 1.0, the Weibull distribution is identical to the exponential distribution. The Weibull location parameter lets you set up an exponential distribution to start at a location other than 0.0. When the shape parameter is less than 1.0, the Weibull distribution becomes a steeply declining curve. A manufacturer might find this effect useful in describing part failures during a burn-in period.

The mathematical constructs for the Weibull distribution are as follows:

$$f(x) = \frac{\alpha}{\beta}\left[\frac{x}{\beta}\right]^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}}$$

$$Mean = \beta\Gamma(1+\alpha^{-1})$$

$$Standard\ Deviation = \beta^2\left[\Gamma(1+2\alpha^{-1}) - \Gamma^2(1+\alpha^{-1})\right]$$

$$Skewness = \frac{2\Gamma^3(1+\beta^{-1}) - 3\Gamma(1+\beta^{-1})\Gamma(1+2\beta^{-1}) + \Gamma(1+3\beta^{-1})}{\left[\Gamma(1+2\beta^{-1}) - \Gamma^2(1+\beta^{-1})\right]^{3/2}}$$

Excess Kurtosis =

$$\frac{-6\Gamma^4(1+\beta^{-1}) + 12\Gamma^2(1+\beta^{-1})\Gamma(1+2\beta^{-1}) - 3\Gamma^2(1+2\beta^{-1}) - 4\Gamma(1+\beta^{-1})\Gamma(1+3\beta^{-1}) + \Gamma(1+4\beta^{-1})}{\left[\Gamma(1+2\beta^{-1}) - \Gamma^2(1+\beta^{-1})\right]^2}$$

Shape (α) and central location scale (β) are the distributional parameters, and Γ is the Gamma function.

Input requirements: Shape Alpha ≥ 0.05. Scale Beta > 0 and can be any positive value.

Weibull 3 Distribution

The Weibull 3 distribution uses the same constructs as the original Weibull distribution but adds a Location, or Shift, parameter. The Weibull distribution starts from a minimum value of 0, whereas this Weibull 3, or Shifted Weibull, distribution shifts the starting location to any other value.

Alpha, Beta, and Location or Shift are the distributional parameters.

Input requirements:

Alpha (Shape) ≥ 0.05.

Beta (Central Location Scale) > 0 and can be any positive value.

Location can be any positive or negative value including zero.

3

3. FORECASTING

Forecasting is the act of predicting the future. It can be based on historical data or speculation about the future when no history exists. When historical data exist, a quantitative or statistical approach is best, but if no historical data exist, then potentially a qualitative or judgmental approach is usually the only recourse. Figure 3.1 lists the most common methodologies for forecasting.

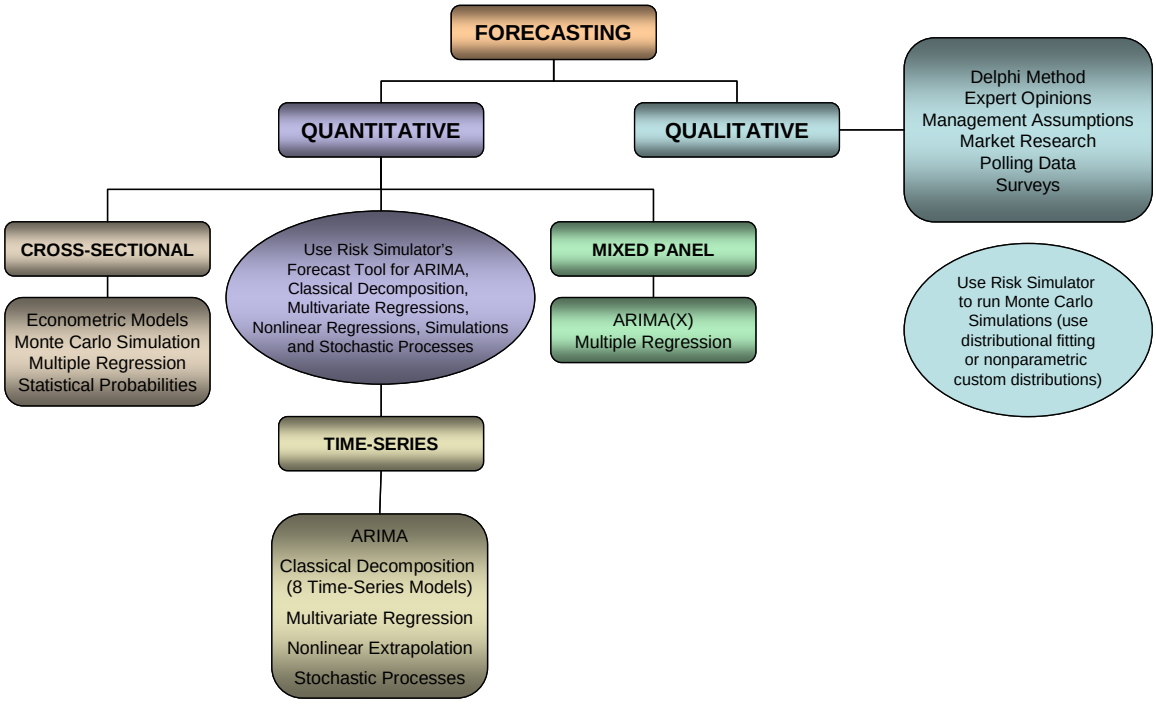


Figure 3.1 – Forecasting Methods

3.1 Different Types of Forecasting Techniques

Generally, forecasting can be divided into quantitative and qualitative approaches. Qualitative forecasting is used when little to no reliable historical, contemporaneous, or comparable data are available. Several qualitative methods exist such as the Delphi, or expert opinion, approach (a consensus-building forecast by field experts, marketing experts, or internal staff members), management assumptions (target growth rates set by senior management), and market research or external data or polling and surveys (data obtained from third-party sources, industry and sector indexes, or active market research). These estimates can be either single-point estimates (an average consensus) or a set of forecast values (a distribution of forecasts). The latter can be entered into Risk Simulator as a custom distribution and the resulting forecasts can be simulated, that is, a nonparametric simulation using the estimated data points themselves as the distribution.

On the quantitative side of forecasting, the available data or data that need to be forecasted can be divided into time-series (values that have a time element to them, such as revenues at different years, inflation rates, interest rates, market share, failure rates), cross-sectional (values that are time-independent, such as the grade point average of sophomore students across the nation in a particular year, given each student's levels of SAT scores, IQ, and number of alcoholic beverages consumed per week), or mixed panel (mixture between time-series and panel data, e.g., predicting sales over the next 10 years given budgeted marketing expenses and market share projections, which means that the sales data is time series but exogenous variables, such as marketing expenses and market share, exist to help to model the forecast predictions).

The Risk Simulator software provides the user several forecasting methodologies:

1. ARIMA (Autoregressive Integrated Moving Average)
2. Auto ARIMA
3. Auto Econometrics
4. Basic Econometrics
5. Combinatorial Fuzzy Logic
6. Cubic Spline Curves
7. Custom Distributions
8. GARCH (Generalized Autoregressive Conditional Heteroskedasticity)
9. J Curve
10. Markov Chain
11. Maximum Likelihood (Logit, Probit, Tobit)
12. Multivariate Regression
13. Neural Network Forecasts
14. Nonlinear Extrapolation
15. S Curve
16. Stochastic Processes
17. Time-Series Analysis and Decomposition
18. Trendlines

The analytical details of each forecasting method fall outside the purview of this user manual. For more details, please review *Modeling Risk*: *Applying Monte Carlo Simulation, Real Options Analysis, Stochastic Forecasting, and Portfolio Optimization*, *Second Edition,* by Dr. Johnathan Mun (Wiley Finance, 2010), who is also the creator of the Risk Simulator software. Nonetheless, the following illustrates some of the more common approaches and several quick getting started

examples in using the software. More detailed descriptions and example models of each of these techniques are found throughout this chapter and the next. All other forecasting approaches are fairly easy to apply within Risk Simulator.

ARIMA

Autoregressive integrated moving average (ARIMA, also known as Box-Jenkins ARIMA) is an advanced econometric modeling technique. ARIMA looks at historical time-series data and performs backfitting optimization routines to account for historical autocorrelation (the relationship of one value versus another in time) and the stability of the data to correct for the nonstationary characteristics of the data, and this predictive model learns over time by correcting its forecasting errors. Advanced knowledge in econometrics is typically required to build good predictive models using this approach.

Auto ARIMA

The Auto ARIMA module automates some of the traditional ARIMA modeling by automatically testing multiple permutations of model specifications and returns the best-fitting model. Running the Auto ARIMA is similar to regular ARIMA forecasts. The difference being that the P, D, Q inputs are no longer required and different combinations of these inputs are automatically run and compared.

Basic Econometrics

Econometrics refers to a branch of business analytics, modeling, and forecasting techniques for modeling the behavior of or forecasting certain business, economic, finance, physics, manufacturing, operations, and any other variables. Running the Basic Econometrics models are similar to regular regression analysis except that the dependent and independent variables are allowed to be modified before a regression is run.

Auto Econometrics

Similar to basic econometrics, but Auto Econometrics allows thousands of linear, nonlinear, interacting, lagged, and mixed variables to be automatically run on your data to determine the best-fitting econometric model that describes the behavior of the dependent variable. It is useful for modeling the effects of the variables and for forecasting future outcomes, while not requiring the analyst to be an expert econometrician.

Combinatorial Fuzzy Logic

In contrast, the term fuzzy logic is derived from fuzzy set theory to deal with reasoning that is approximate rather than accurate—as opposed to crisp logic, where binary sets have binary logic, fuzzy logic variables may have a truth value that ranges between 0 and 1 and is not constrained to the two truth values of classic propositional logic. This fuzzy weighting schema is used together with a combinatorial method to yield time-series forecast results.

Cubic Spline Curves

Sometimes there are missing values in a time-series data set. For instance, interest rates for years 1 to 3 may exist, followed by years 5 to 8, and then year 10. Spline curves can be used to interpolate the missing years' interest rate values based on the data that exist. Spline curves can also be used to forecast or extrapolate values of future time periods beyond the time period of available data. The data can be linear or nonlinear.

Custom Distributions

Using Risk Simulator, expert opinions can be collected and a customized distribution can be generated. This forecasting technique comes in handy when the data set is small or the goodness of fit is bad when applied to a distributional fitting routine.

GARCH

The generalized autoregressive conditional heteroskedasticity (GARCH) model is used to model historical and forecast future volatility levels of a marketable security (e.g., stock prices, commodity prices, and oil prices). The data set has to be a time series of raw price levels. GARCH will first convert the prices into relative returns and then run an internal optimization to fit the historical data to a mean-reverting volatility term structure, while assuming that the volatility is heteroskedastic in nature (changes over time according to some econometric characteristics). Several variations of this methodology are available in Risk Simulator, including EGARCH, EGARCH-T, GARCH-M, GJR-GARCH, GJR-GARCH-T, IGARCH, and T-GARCH.

J Curve

The J curve, or exponential growth curve, is where the growth of the next period depends on the current period's level and the increase is exponential. This means that over time, the values will increase significantly from one period to another. This model is typically used in forecasting biological growth and chemical reactions over time.

Markov Chain

A Markov chain exists when the probability of a future state depends on a previous state and when linked together form a chain that reverts to a long-run steady state level. This approach is typically used to forecast the market share of two competitors. The required inputs are the starting probability of a customer in the first store (the first state) will return to the same store in the next period versus the probability of switching to a competitor's store in the next state.

Maximum Likelihood on Logit, Probit, and Tobit

Maximum likelihood estimation (MLE) is used to forecast the probability of something occurring given some independent variables. For instance, MLE is used to predict if a credit line or debt will default given the obligor's characteristics (30 years old, single, salary of $100,000 per year, and having a total credit card debt of $10,000); or the probability a patient will have lung cancer if the person is a male between the ages of 50 and 60, smokes 5 packs of cigarettes per month, and so forth. In these circumstances, the dependent variable is limited (i.e., limited to being binary 1 and 0 for default/die and no default/live, or limited to integer values like 1, 2, 3,etc.), and the desired outcome of the model is to predict the probability of an event occurring. Traditional regression analysis will not work in these situations (the predicted probability is usually less than zero or greater than one, and many of the required regression assumptions are violated, such as independence and normality of the errors, and the errors will be fairly large).

Multivariate Regression

Multivariate regression is used to model the relationship structure and characteristics of a certain dependent variable as it depends on other independent exogenous variables. Using the modeled relationship, we can forecast the future values of the dependent variable. The accuracy and goodness of fit for this model can also be determined. Linear and nonlinear models can be fitted in the multiple regression analysis.

Neural Network Forecast

The term Neural Network is often used to refer to a network or circuit of biological neurons, while modern usage of the term often refers to artificial neural networks comprising artificial neurons, or nodes, recreated in a software environment. Such networks attempt to mimic the neurons in the human brain in ways of thinking and identifying patterns and, in our situation, identifying patterns for the purposes of forecasting time-series data.

Nonlinear Extrapolation

The underlying structure of the data to be forecasted is assumed to be nonlinear over time. For instance, a data set such as 1, 4, 9, 16, 25 is considered to be nonlinear (these data points are from a squared function).

S Curve

The S curve or logistic growth curve starts off like a J curve, with exponential growth rates. Over time, the environment becomes saturated (e.g., market saturation, competition, overcrowding), the growth slows, and the forecast value eventually ends up at a saturation or maximum level. This model is typically used in forecasting market share or sales growth of a new product from market introduction until maturity and decline, population dynamics, and other naturally occurring phenomenon.

Stochastic Processes

Sometimes variables cannot be readily predicted using traditional means, and these variables are said to be stochastic. Nonetheless, most financial, economic, and naturally occurring phenomena (e.g., motion of molecules through the air) follow a known mathematical law or relationship. Although the resulting values are uncertain, the underlying mathematical structure is known and can be simulated using Monte Carlo risk simulation. The processes supported in Risk Simulator include Brownian motion random walk, mean-reversion, jump-diffusion, and mixed processes, useful for forecasting nonstationary time-series variables.

Time-Series Analysis and Decomposition

In well-behaved time-series data (typical examples include sales revenues and cost structures of large corporations), the values tend to have up to three elements: a base value, trend, and seasonality. Time-series analysis uses these historical data and decomposes them into these three elements, and recomposes them into future forecasts. In other words, this forecasting method, like some of the others described, first performs a back-fitting (backcast) of historical data before it provides estimates of future values (forecasts).

Trendlines

Trendlines can be used to determine if a set of time-series data follows any appreciable trend. Trends can be linear or nonlinear (such as exponential, logarithmic, moving average, power, polynomial, or power).

3.2 Running the Forecasting Tool in Risk Simulator

In general, to create forecasts, several quick steps are required:

- Start Excel and enter in or open your existing historical data.
- Select the data, and click on *Simulation* and select *Forecasting*.
- Select the relevant sections (ARIMA, Multivariate Regression, Nonlinear Extrapolation, Stochastic Forecasting, Time-Series Analysis) and enter the relevant inputs.

Figure 3.2 illustrates the *Forecasting* tool and the various methodologies and the following provides a quick review of the selected methodology and several quick getting started examples in using the software. The example file can be found either on the start menu at *Start | Real Options Valuation | Risk Simulator | Examples* or accessed directly through *Risk Simulator | Example Models*.

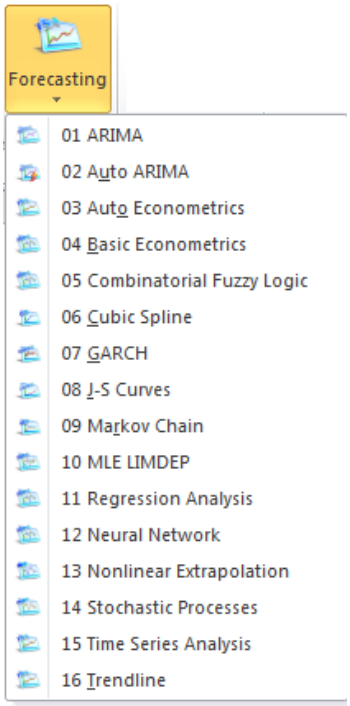


Figure 3.2 – *Risk Simulator's* Forecasting Methods

3.3 Time-Series Analysis

Theory

Figure 3.3 lists the eight most common time-series models, segregated by seasonality and trend. For instance, if the data variable has no trend or seasonality, then a single moving-average model or a single exponential-smoothing model would suffice. However, if seasonality exists but no discernible trend is present, either a seasonal additive or seasonal multiplicative model would be better, and so forth.

	No Seasonality	With Seasonality
No Trend	*Single Moving Average*	*Seasonal Additive*
	Single Exponential Smoothing	*Seasonal Multiplicative*
With Trend	*Double Moving Average*	*Holt-Winter's Additive*
	Double Exponential Smoothing	*Holt-Winter's Multiplicative*

Figure 3.3 – The Eight Most Common Time-Series Methods

Procedure

- Start Excel and open your historical data if required (the example below uses the Time-Series Forecasting file in the examples folder).
- *Select the historical data* (data should be listed in a single column).
- Select *Risk Simulator | Forecasting | Time-Series Analysis*.
- Choose the model to apply, enter the relevant assumptions (Figure 3.4), and click *OK*

Results Interpretation

Figure 3.5 illustrates the sample results generated by using the *Forecasting* tool and a Holt-Winter's multiplicative model. The model-fitting and forecast chart indicates that the trend and seasonality are picked up nicely by the Holt-Winter's multiplicative model. The time-series analysis report provides the relevant optimized alpha, beta, and gamma parameters; the error measurements; fitted data; forecast values; and fitted-forecast graph. The parameters are simply for reference. Alpha captures the memory effect of the base level changes over time, and beta is the trend parameter that measures the strength of the trend, while gamma measures the seasonality strength of the historical data. The analysis decomposes the historical data into these three elements and then recomposes them to forecast the future. The fitted data illustrates the historical data, and it uses the recomposed model and shows how close the forecasts are in the past (a technique called *backcasting*). The forecast values are either single-point estimates or assumptions (if the option to automatically generate assumptions is chosen and if a simulation profile exists). The graph illustrates these historical, fitted, and forecast values. The chart is a powerful communication and visual tool to see how good the forecast model is.

Historical Sales Revenues

Year	Quarter	Period	Sales
2006	1	1	$684.20
2006	2	2	$584.10
2006	3	3	$765.40
2006	4	4	$892.30
2007	1	5	$885.40
2007	2	6	$677.00
2007	3	7	$1,006.60
2007	4	8	$1,122.10
2008	1	9	$1,163.40
2008	2	10	$993.20
2008	3	11	$1,312.50
2008	4	12	$1,545.30
2009	1	13	$1,596.20
2009	2	14	$1,260.40
2009	3	15	$1,735.20
2009	4	16	$2,029.70
2010	1	17	$2,107.80
2010	2	18	$1,650.30
2010	3	19	$2,304.40
2010	4	20	$2,639.40

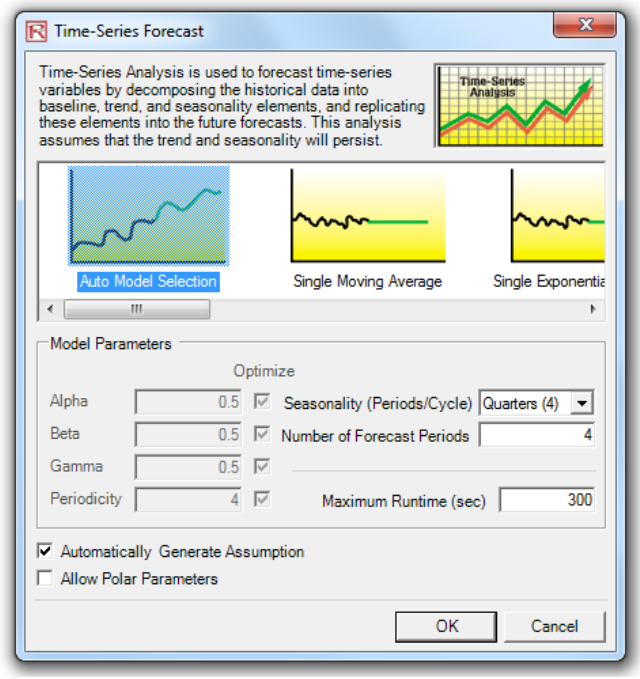


Figure 3.4 – Time-Series Analysis

Notes

This time-series analysis module contains the eight time-series models seen in Figure 3.3. You can choose the specific model to run based on the trend and seasonality criteria or choose the *Auto Model Selection*, which will automatically iterate through all eight methods, optimize the parameters, and find the best-fitting model for your data. Alternatively, if you choose one of the eight models, you can also unselect the *optimize* checkboxes and enter your own alpha, beta, and gamma parameters. Refer to Dr. Johnathan Mun's *Modeling Risk: Applying Monte Carlo Simulation, Real Options Analysis, Forecasting, and Optimization, Second Edition* (Wiley Finance, 2010) for more details on the technical specifications of these parameters. In addition, you would need to enter the relevant seasonality periods if you choose the automatic model selection or any of the seasonal models. The seasonality input has to be a positive integer (e.g., if the data is quarterly, enter 4 as the number of seasons or cycles a year, or enter 12 if monthly data). Next, enter the number of periods to forecast. This value also has to be a positive integer. The maximum runtime is set at 300 seconds. Typically, no changes are required. However, when forecasting with a significant amount of historical data, the analysis might take slightly longer, and if the processing time exceeds this runtime, the process will be terminated. You can also elect to have the forecast automatically generate assumptions. That is, instead of single-point estimates, the forecasts will be assumptions. Finally, the polar parameters option allows you to optimize the alpha, beta, and gamma parameters to include zero and one. Certain forecasting software allows these polar parameters while others do not. Risk Simulator allows you to choose which to use. Typically, there is no need to use polar parameters.

Holt-Winter's Multiplicative

Summary Statistics

Alpha, Beta, Gamma	RMSE	Alpha, Beta, Gamma	RMSE
0.00, 0.00, 0.00	914.824	0.00, 0.00, 0.00	914.824
0.10, 0.10, 0.10	415.322	0.10, 0.10, 0.10	415.322
0.20, 0.20, 0.20	187.202	0.20, 0.20, 0.20	187.202
0.30, 0.30, 0.30	118.795	0.30, 0.30, 0.30	118.795
0.40, 0.40, 0.40	101.794	0.40, 0.40, 0.40	101.794
0.50, 0.50, 0.50	102.143		

The analysis was run with alpha = 0.2429, beta = 1.0000, gamma = 0.7797, and seasonality = 4

Time-Series Analysis Summary

When both seasonality and trend exist, more advanced models are required to decompose the data into their base elements: a base-case level (L) weighted by the alpha parameter; a trend component (b) weighted by the beta parameter; and a seasonality component (S) weighted by the gamma parameter. Several methods exist but the two most common are the Holt-Winters' additive seasonality and Holt-Winters' multiplicative seasonality methods. In the Holt-Winter's additive model, the base case level, seasonality, and trend are added together to obtain the forecast fit.

The best-fitting test for the moving average forecast uses the root mean squared errors (RMSE). The RMSE calculates the square root of the average squared deviations of the fitted values versus the actual data points.

Mean Squared Error (MSE) is an absolute error measure that squares the errors (the difference between the actual historical data and the forecast-fitted data predicted by the model) to keep the positive and negative errors from canceling each other out. This measure also tends to exaggerate large errors by weighting the large errors more heavily than smaller errors by squaring them, which can help when comparing different time-series models. Root Mean Square Error (RMSE) is the square root of MSE and is the most popular error measure, also known as the quadratic loss function. RMSE can be defined as the average of the absolute values of the forecast errors and is highly appropriate when the cost of the forecast errors is proportional to the absolute size of the forecast error. The RMSE is used as the selection criteria for the best-fitting time-series model.

Mean Absolute Percentage Error (MAPE) is a relative error statistic measured as an average percent error of the historical data points and is most appropriate when the cost of the forecast error is more closely related to the percentage error than the numerical size of the error. Finally, an associated measure is the Theil's U statistic, which measures the naivety of the model's forecast. That is, if the Theil's U statistic is less than 1.0, then the forecast method used provides an estimate that is statistically better than guessing.

Period	Actual	Forecast Fit
1	684.20	
2	584.10	
3	765.40	
4	892.30	
5	885.40	684.20
6	677.00	667.55
7	1006.60	935.45
8	1122.10	1198.09
9	1163.40	1112.48
10	993.20	887.95
11	1312.50	1348.38
12	1545.30	1546.53
13	1596.20	1572.44
14	1260.40	1299.20
15	1735.20	1704.77
16	2029.70	1976.23
17	2107.80	2026.01
18	1650.30	1637.28
19	2304.40	2245.93
20	2639.40	2643.09
Forecast 21		2713.69
Forecast 22		2114.79
Forecast 23		2900.42
Forecast 24		3293.81

Error Measurements	
RMSE	71.8132
MSE	5157.1348
MAD	53.4071
MAPE	4.50%
Theil's U	0.3054

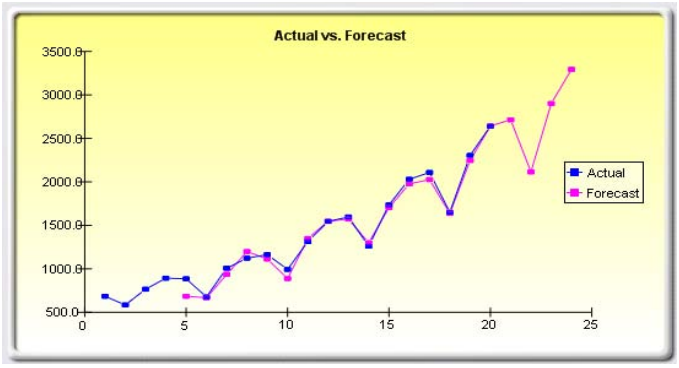


Figure 3.5 – Example Holt-Winter's Forecast Report

3.4 Multivariate Regression

Theory

It is assumed that the user is sufficiently knowledgeable about the fundamentals of regression analysis. The general bivariate linear regression equation takes the form of $Y = \beta_0 + \beta_1 X + \varepsilon$, where β_0 is the intercept, β_1 is the slope, and ε is the error term. It is bivariate as there are only two variables: a *Y,* or dependent, variable and an *X,* or independent, variable, where *X* is also known as the regressor (sometimes a bivariate regression is also known as a univariate regression as there is only a single independent variable *X*). The dependent variable is so named because it *depends* on the independent variable; for example, sales revenue depends on the amount of marketing costs expended on a product's advertising and promotion, making the dependent variable sales and the independent variable marketing costs. An example of a bivariate regression is seen as simply inserting the best-fitting line through a set of data points in a two-dimensional plane as seen on the left panel in Figure 3.6. In other cases, a multivariate regression can be performed, where there are multiple, or *n* number of, independent *X* variables, where the general regression equation will now take the form of $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 ... + \beta_n X_n + \varepsilon$. In this case, the best-fitting line will be within an $n + 1$ dimensional plane.

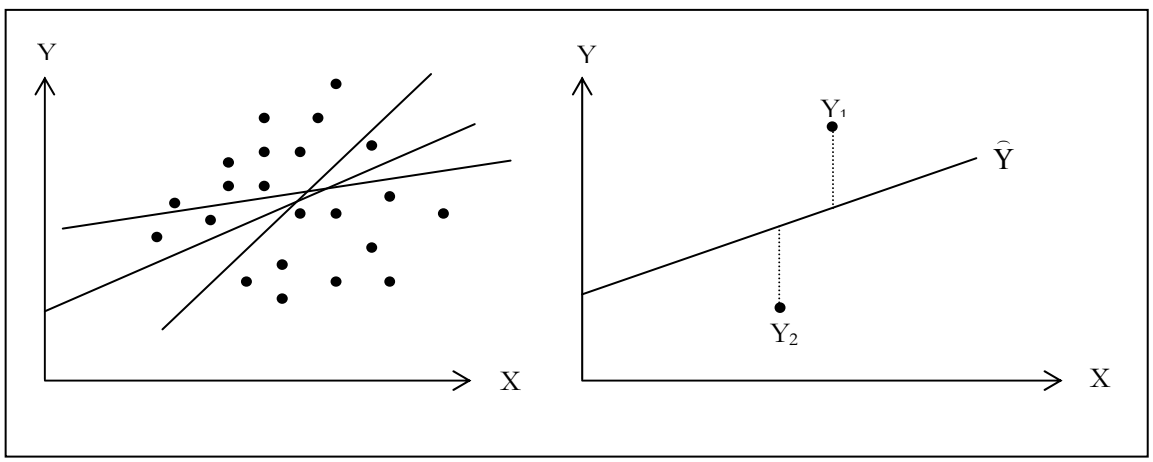


Figure 3.6 – Bivariate Regression

However, fitting a line through a set of data points in a scatter plot as in Figure 3.6 may result in numerous possible lines. The best-fitting line is defined as the single unique line that minimizes the total vertical errors, that is, the sum of the absolute distances between the actual data points (Y_i) and the estimated line ($\hat{Y}$) as shown on the right panel of Figure 3.6. To find the best-fitting line that minimizes the errors, a more sophisticated approach is required, that is, regression analysis. Regression analysis, therefore, finds the unique best-fitting line by requiring that the total errors be minimized, or by calculating

$$Min \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

where only one unique line minimizes this sum of squared errors. The errors (vertical distance between the actual data and the predicted line) are squared to avoid the negative errors

canceling out the positive errors. Solving this minimization problem with respect to the slope and intercept requires calculating a first derivative and setting them equal to zero:

$$\frac{d}{d\beta_0}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=0 \text{ and } \frac{d}{d\beta_1}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=0$$

which yields the bivariate regression's least squares equations:

$$\beta_1=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}=\frac{\sum_{i=1}^{n}X_iY_i-\frac{\sum_{i=1}^{n}X_i\sum_{i=1}^{n}Y_i}{n}}{\sum_{i=1}^{n}X_i^2-\frac{\left(\sum_{i=1}^{n}X_i\right)^2}{n}}$$

$$\beta_0=\bar{Y}-\beta_1\bar{X}$$

For multivariate regression, the analogy is expanded to account for multiple independent variables, where $Y_i=\beta_1+\beta_2X_{2,i}+\beta_3X_{3,i}+\varepsilon_i$ and the estimated slopes can be calculated by:

$$\hat{\beta}_2=\frac{\sum Y_iX_{2,i}\sum X_{3,i}^2-\sum Y_iX_{3,i}\sum X_{2,i}X_{3,i}}{\sum X_{2,i}^2\sum X_{3,i}^2-\left(\sum X_{2,i}X_{3,i}\right)^2}$$

$$\hat{\beta}_3=\frac{\sum Y_iX_{3,i}\sum X_{2,i}^2-\sum Y_iX_{2,i}\sum X_{2,i}X_{3,i}}{\sum X_{2,i}^2\sum X_{3,i}^2-\left(\sum X_{2,i}X_{3,i}\right)^2}$$

In running multivariate regressions, great care has to be taken to set up and interpret the results. For instance, a good understanding of econometric modeling is required (e.g., identifying regression pitfalls such as structural breaks, multicollinearity, heteroskedasticity, autocorrelation, specification tests, nonlinearities, etc.) before a proper model can be constructed. See *Modeling Risk: Applying Monte Carlo Simulation, Real Options Analysis, Forecasting, and Optimization, Second Edition* (Wiley Finance, 2010) by Dr. Johnathan Mun for more detailed analysis and discussion of multivariate regression as well as how to identify these regression pitfalls.

Procedure

- Start Excel and open your historical data if required (the illustration below uses the file Multiple Regression in the examples folder).
- Check to make sure that the data is arranged in columns, select the entire data area including the variable name, and select *Risk Simulator | Forecasting | Multiple Regression.*
- Select the dependent variable and check the relevant options (lags, stepwise regression, nonlinear regression, etc.), and click *OK.*

Results Interpretation

Figure 3.8 illustrates a sample multivariate regression result report. The report comes complete with all the regression results, analysis of variance results, fitted chart, and hypothesis test results. The technical details of interpreting these results are beyond the scope of this user manual. See *Modeling Risk: Applying Monte Carlo Simulation, Real Options Analysis, Forecasting, and Optimization, Second Edition* (Wiley Finance, 2010) by Dr. Johnathan Mun for more detailed analysis and discussion of multivariate regression as well as the interpretation of regression reports.

Multivariate Regression

Y	X1	X2	X3	X4	X5
521	18308	185	4.041	79.6	7.2
367	1148	600	0.55	1	8.5
443	18068	372	3.665	32.3	5.7
365	7729	142	2.351	45.1	7.3
614	100484	432	29.76	190.8	7.5
385	16728	290	3.294	31.8	5
286	14630	346	3.287	678.4	6.7
397	4008	328	0.666	340.8	6.2
764	38927	354	12.938	239.6	7.3
427	22322	266	6.478	111.9	5
153	3711	320	1.108	172.5	2.8
231	3136	197	1.007	12.2	6.1
524	50508	266	11.431	205.6	7.1
328	28886	173	5.544	154.6	5.9
240	16996	190	2.777	49.7	4.6
286	13035	239	2.478	30.3	4.4
285	12973	190	3.685	92.8	7.4
569	16309	241	4.22	96.9	7.1
96	5227	189	1.228	39.8	7.5
498	19235	358	4.781	489.2	5.9
481	44487	315	6.016	767.6	9
468	44213	303	9.295	163.6	9.2
177	23619	228	4.375	55	5.1
198	9106	134	2.573	54.9	8.6
458	24917	189	5.117	74.3	6.6
108	3872	196	0.799	5.5	6.9
246	8945	183	1.578	20.5	2.7
291	2373	417	1.202	10.9	5.5
68	7128	233	1.109	123.7	7.2
311	23624	349	7.73	1042	6.6
606	5242	284	1.515	12.5	6.9
512	92629	499	17.99	381	7.2
426	28795	231	6.629	136.1	5.8
47	4487	143	0.639	9.3	4.1
265	48799	249	10.847	264.9	6.4

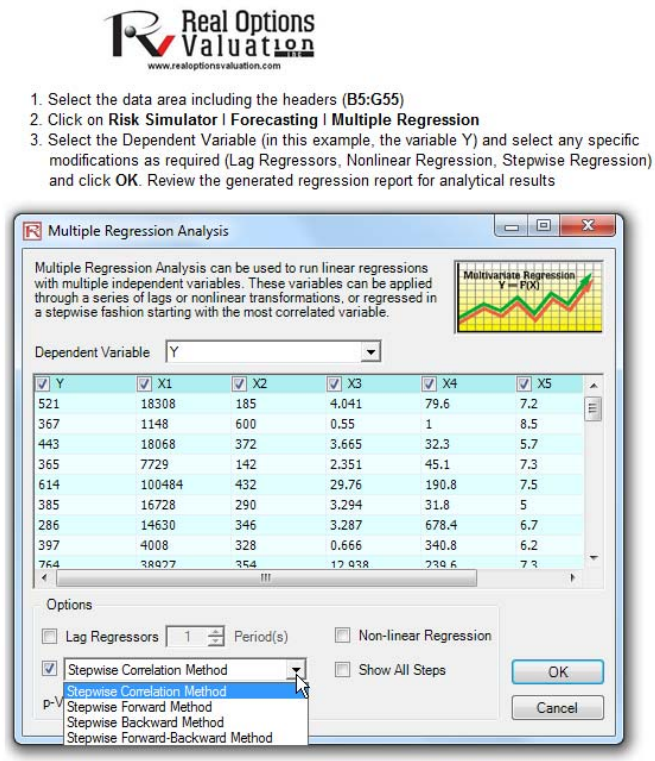


Figure 3.7 – Running a Multivariate Regression

Regression Analysis Report

Regression Statistics

R-Squared (Coefficient of Determination)	0.3272
Adjusted R-Squared	0.2508
Multiple R (Multiple Correlation Coefficient)	0.5720
Standard Error of the Estimates (SEy)	149.6720
Number of Observations	50

The R-Squared or Coefficient of Determination indicates that 0.33 of the variation in the dependent variable can be explained and accounted for by the independent variables in this regression analysis. However, in a multiple regression, the Adjusted R-Squared takes into account the existence of additional independent variables or regressors and adjusts this R-Squared value to a more accurate view of the regression's explanatory power. Hence, only 0.25 of the variation in the dependent variable can be explained by the regressors.

The Multiple Correlation Coefficient (Multiple R) measures the correlation between the actual dependent variable (Y) and the estimated or fitted (Y) based on the regression equation. This is also the square root of the Coefficient of Determination (R-Squared).

The Standard Error of the Estimates (SEy) describes the dispersion of data points above and below the regression line or plane. This value is used as part of the calculation to obtain the confidence interval of the estimates later.

Regression Results

	Intercept	X1	X2	X3	X4	X5
Coefficients	57.9555	-0.0035	0.4644	25.2377	-0.0086	16.5579
Standard Error	108.7901	0.0035	0.2535	14.1172	0.1016	14.7996
t-Statistic	0.5327	-1.0066	1.8316	1.7877	-0.0843	1.1188
p-Value	0.5969	**0.3197**	**0.0738**	**0.0807**	**0.9332**	**0.2693**
Lower 5%	-161.2966	-0.0106	-0.0466	-3.2137	-0.2132	-13.2687
Upper 95%	277.2076	0.0036	0.9753	53.6891	0.1961	46.3845

Degrees of Freedom		**Hypothesis Test**	
Degrees of Freedom for Regression	5	Critical t-Statistic (99% confidence with df of 44)	2.6923
Degrees of Freedom for Residual	44	Critical t-Statistic (95% confidence with df of 44)	2.0154
Total Degrees of Freedom	49	Critical t-Statistic (90% confidence with df of 44)	1.6802

The Coefficients provide the estimated regression intercept and slopes. For instance, the coefficients are estimates of the true; population b values in the following regression equation Y = b0 + b1X1 + b2X2 + ... + bnXn. The Standard Error measures how accurate the predicted Coefficients are, and the t-Statistics are the ratios of each predicted Coefficient to its Standard Error.

The t-Statistic is used in hypothesis testing, where we set the null hypothesis (Ho) such that the real mean of the Coefficient = 0, and the alternate hypothesis (Ha) such that the real mean of the Coefficient is not equal to 0. A t-test is is performed and the calculated t-Statistic is compared to the critical values at the relevant Degrees of Freedom for Residual. The t-test is very important as it calculates if each of the coefficients is statistically significant in the presence of the other regressors. This means that the t-test statistically verifies whether a regressor or independent variable should remain in the regression or it should be dropped.

The Coefficient is statistically significant if its calculated t-Statistic exceeds the Critical t-Statistic at the relevant degrees of freedom (df). The three main confidence levels used to test for significance are 90%, 95% and 99%. If a Coefficient's t-Statistic exceeds the Critical level, it is considered statistically significant. Alternatively, the p-Value calculates each t-Statistic's probability of occurrence, which means that the smaller the p-Value, the more significant the Coefficient. The usual significant levels for the p-Value are 0.01, 0.05, and 0.10, corresponding to the 99%, 95%, and 90% confidence levels.

The Coefficients with their p-Values highlighted in blue indicate that they are statistically significant at the 90% confidence or 0.10 alpha level, while those highlighted in red indicate that they are not statistically significant at any other alpha levels.

Analysis of Variance

	Sums of Squares	Mean of Squares	F-Statistic	p-Value	**Hypothesis Test**	
Regression	479388.49	95877.70	4.28	**0.0029**	Critical F-statistic (99% confidence with df of 5 and 44)	3.4651
Residual	985675.19	22401.71			Critical F-statistic (95% confidence with df of 5 and 44)	2.4270
Total	1465063.68				Critical F-statistic (90% confidence with df of 5 and 44)	1.9828

The Analysis of Variance (ANOVA) table provides an F-test of the regression model's overall statistical significance. Instead of looking at individual regressors as in the t-test, the F-test looks at all the estimated Coefficients' statistical properties. The F-Statistic is calculated as the ratio of the Regression's Mean of Squares to the Residual's Mean of Squares. The numerator measures how much of the regression is explained, while the denominator measures how much is unexplained. Hence, the larger the F-Statistic, the more significant the model. The corresponding p-Value is calculated to test the null hypothesis (Ho) where all the Coefficients are simultaneously equal to zero, versus the alternate hypothesis (Ha) that they are all simultaneously different from zero, indicating a significant overall regression model. If the p-Value is smaller than the 0.01, 0.05, or 0.10 alpha significance, then the regression is significant. The same approach can be applied to the F-Statistic by comparing the calculated F-Statistic with the critical F values at various significance levels.

Forecasting

Period	Actual (Y)	Forecast (F)	Error (E)
1	521.0000	299.5124	221.4876
2	367.0000	487.1243	(120.1243)
3	443.0000	353.2789	89.7211
4	365.0000	276.3296	88.6704
5	614.0000	776.1336	(162.1336)
6	385.0000	298.9993	86.0007
7	286.0000	354.8718	(68.8718)
8	397.0000	312.6155	84.3845
9	764.0000	529.7550	234.2450
10	427.0000	347.7034	79.2966
11	153.0000	266.2526	(113.2526)
12	231.0000	264.6375	(33.6375)
13	524.0000	406.8009	117.1991
14	328.0000	272.2226	55.7774
15	240.0000	231.7882	8.2118
16	286.0000	257.8862	28.1138
17	285.0000	314.9521	(29.9521)
18	569.0000	335.3140	233.6860
19	96.0000	282.0356	(186.0356)
20	498.0000	370.2062	127.7938
21	481.0000	340.8742	140.1258
22	468.0000	427.5118	40.4882
23	177.0000	274.5298	(97.5298)
24	198.0000	294.7795	(96.7795)
25	458.0000	295.2180	162.7820

RMSE: 140.4048

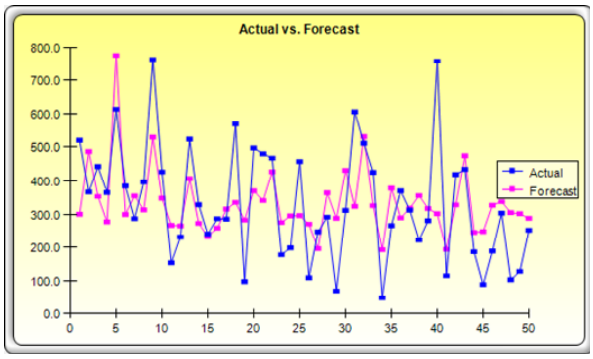


Figure 3.8 – Multivariate Regression Results

3.5 Stochastic Forecasting

Theory

A stochastic process is nothing but a mathematically defined equation that can create a series of outcomes over time, outcomes that are not deterministic in nature, that is, an equation or process that does not follow any simple discernible rule such as price will increase *X* percent every year or revenues will increase by this factor of *X* plus *Y* percent. A stochastic process is by definition nondeterministic, and one can plug numbers into a stochastic process equation and obtain different results every time. For instance, the path of a stock price is stochastic in nature, and one cannot reliably predict the stock price path with any certainty. However, the price evolution over time is enveloped in a process that generates these prices. The process is fixed and predetermined, but the outcomes are not. Hence, by stochastic simulation, we create multiple pathways of prices, obtain a statistical sampling of these simulations, and make inferences on the potential pathways that the actual price may undertake given the nature and parameters of the stochastic process used to generate the time series. Three basic stochastic processes are included in Risk Simulator's *Forecasting* tool, including geometric Brownian motion or random walk, which is the most common and prevalently used process due to its simplicity and wide-ranging applications. The other two stochastic processes are the mean-reversion process and the jump-diffusion process.

The interesting thing about stochastic process simulation is that historical data are not necessarily required. That is, the model does not have to fit any sets of historical data. Simply compute the expected returns and the volatility of the historical data or estimate them using comparable external data or make assumptions about these values. See *Modeling Risk: Applying Monte Carlo Simulation, Real Options Analysis, Forecasting, and Optimization, Second Edition* (Wiley Finance, 2010) by Dr. Johnathan Mun for more details on how each of the inputs are computed (e.g., mean-reversion rate, jump probabilities, volatility, etc.).

Procedure

- Start the module by selecting *Risk Simulator | Forecasting | Stochastic Processes*.
- Select the desired process, enter the required inputs, click on *Update Chart* a few times to make sure the process is behaving the way you expect it to, and click *OK* (Figure 3.9).

Results Interpretation

Figure 3.10 shows the results of a sample stochastic process. The chart shows a sample set of the iterations while the report explains the basics of stochastic processes. In addition, the forecast values (mean and standard deviation) for each time period are provided. Using these values, you can decide which time period is relevant to your analysis and set assumptions based on these mean and standard deviation values using the normal distribution. These assumptions can then be simulated in your own custom model.

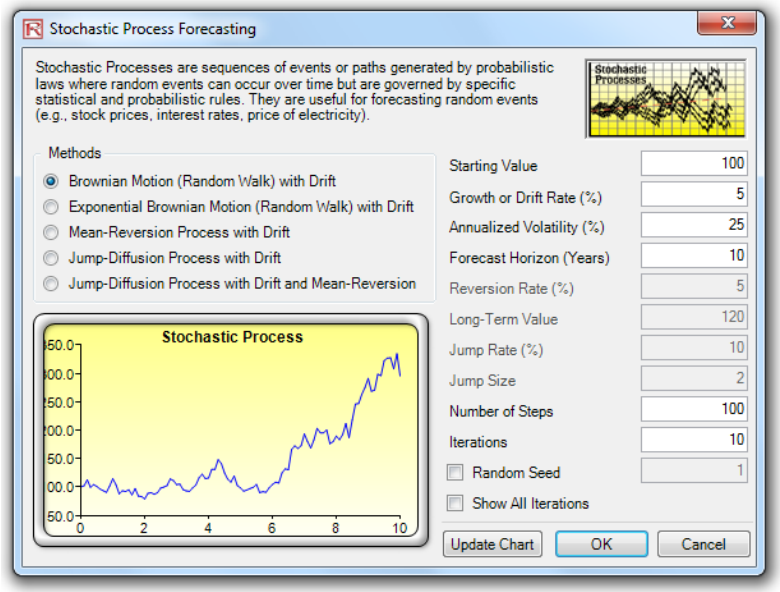


Figure 3.9 – Stochastic Process Forecasting

Stochastic Process Forecasting

Statistical Summary

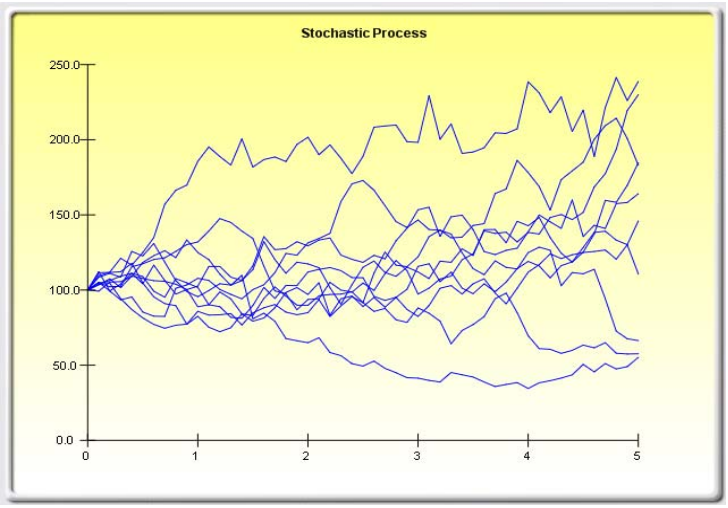
A stochastic process is a sequence of events or paths generated by probabilistic laws. That is, random events can occur over time but are governed by specific statistical and probabilistic rules. The main stochastic processes include Random Walk or Brownian Motion, Mean-Reversion, and Jump-Diffusion. These processes can be used to forecast a multitude of variables that seemingly follow random trends but yet are restricted by probabilistic laws.

The Random Walk Brownian Motion process can be used to forecast stock prices, prices of commodities, and other stochastic time-series data given a drift or growth rate and a volatility around the drift path. The Mean-Reversion process can be used to reduce the fluctuations of the Random Walk process by allowing the path to target a long-term value, making it useful for forecasting time-series variables that have a long-term rate such as interest rates and inflation rates (these are long-term target rates by regulatory authorities or the market). The Jump-Diffusion process is useful for forecasting time-series data when the variable can occasionally exhibit random jumps, such as oil prices or price of electricity (discrete exogenous event shocks can make prices jump up or down). Finally, these three stochastic processes can be mixed and matched as required.

The results on the right indicate the mean and standard deviation of all the iterations generated at each time step. If the Show All Iterations option is selected, each iteration pathway will be shown in a separate worksheet. The graph generated below shows a sample set of the iteration pathways.

Stochastic Process: Brownian Motion (Random Walk) with Drift

Start Value	*100*	*Steps*	*50.00*	*Jump Rate*	*N/A*
Drift Rate	*5.00%*	*Iterations*	*10.00*	*Jump Size*	*N/A*
Volatility	*25.00%*	*Reversion Rate*	*N/A*	*Random Seed*	*1720050445*
Horizon	*5*	*Long-Term Value*	*N/A*		



Time	Mean	Stdev
0.0000	100.00	0.00
0.1000	106.32	4.05
0.2000	105.92	4.70
0.3000	105.23	8.23
0.4000	109.84	11.18
0.5000	107.57	14.67
0.6000	108.63	19.79
0.7000	107.85	24.18
0.8000	109.61	24.46
0.9000	109.57	27.99
1.0000	110.74	30.81
1.1000	111.53	35.05
1.2000	111.07	34.10
1.3000	107.52	32.85
1.4000	108.26	37.38
1.5000	106.36	32.19
1.6000	112.42	32.16
1.7000	110.08	31.24
1.8000	109.64	31.87
1.9000	110.18	36.43
2.0000	112.23	37.63
2.1000	114.32	33.10
2.2000	111.14	38.42
2.3000	111.03	37.69
2.4000	112.04	37.23
2.5000	112.98	40.84
2.6000	115.74	43.69
2.7000	115.11	43.64
2.8000	114.87	43.70
2.9000	113.28	42.25
3.0000	115.72	43.43
3.1000	120.05	50.48
3.2000	116.69	42.61
3.3000	118.31	45.57
3.4000	116.35	40.82
3.5000	115.71	40.33
3.6000	118.69	41.45
3.7000	121.66	45.34
3.8000	121.40	45.03
3.9000	125.19	48.19
4.0000	129.65	55.44
4.1000	129.61	53.82
4.2000	125.86	49.68
4.3000	125.70	53.79
4.4000	126.72	49.70
4.5000	129.52	50.28
4.6000	132.28	49.70
4.7000	138.47	56.77
4.8000	139.69	66.32
4.9000	140.85	65.95
5.0000	143.61	68.65

Figure 3.10 – Stochastic Forecast Result

3.6 Nonlinear Extrapolation

Theory

Extrapolation involves making statistical projections by using historical trends that are projected for a specified period of time into the future. It is only used for time-series forecasts. For cross-sectional or mixed panel data (time-series with cross-sectional data), multivariate regression is more appropriate. Extrapolation is useful when major changes are not expected, that is, causal factors are expected to remain constant or when the causal factors of a situation are not clearly understood. It also helps discourage introduction of personal biases into the process. Extrapolation is fairly reliable, relatively simple, and inexpensive. However, extrapolation, which assumes that recent and historical trends will continue, produces large forecast errors if discontinuities occur within the projected time period. That is, pure extrapolation of time series assumes that all we need to know is contained in the historical values of the series that is being forecasted. If we assume that past behavior is a good predictor of future behavior, extrapolation is appealing. This makes it a useful approach when all that is needed are many short-term forecasts.

This methodology estimates the $f(x)$ function for any arbitrary x value by interpolating a smooth nonlinear curve through all the x values and, using this smooth curve, extrapolates future x values beyond the historical data set. The methodology employs either the polynomial functional form or the rational functional form (a ratio of two polynomials). Typically, a polynomial functional form is sufficient for well-behaved data, however, rational functional forms are sometimes more accurate (especially with polar functions, i.e., functions with denominators approaching zero).

Procedure

- Start Excel and open your historical data if required (the illustration shown next uses the file Nonlinear Extrapolation from the examples folder).
- *Select the time-series data* and select *Risk Simulator | Forecasting | Nonlinear Extrapolation*.
- Select the extrapolation type (automatic selection, polynomial function, or rational function) and enter the number of forecast period desired (Figure 3.11), and click *OK*.

Results Interpretation

The results report shown in Figure 3.12 shows the extrapolated forecast values, the error measurements, and the graphical representation of the extrapolation results. The error measurements should be used to check the validity of the forecast and are especially important when used to compare the forecast quality and accuracy of extrapolation versus time-series analysis.

Notes

When the historical data is smooth and follows some nonlinear patterns and curves, extrapolation is better than time-series analysis. However, when the data patterns follow seasonal cycles and a trend, time-series analysis will provide better results.

Note that Nonlinear Extrapolation involves making statistical projections by using historical trends that are projected for a specified period of time into the future. It is only used for time-series forecasts. Extrapolation is fairly reliable, relatively simple, and inexpensive. However, extrapolation, which assumes that recent and historical trends will continue, produces large forecast errors if discontinuities occur within the projected time period.

Historical Sales Revenues
Polynomial Growth Rates

Year	Month	Period	Sales
2010	1	1	$1.00
2010	2	2	$6.73
2010	3	3	$20.52
2010	4	4	$45.25
2010	5	5	$83.59
2010	6	6	$138.01
2010	7	7	$210.87
2010	8	8	$304.44
2010	9	9	$420.89
2010	10	10	$562.34
2010	11	11	$730.85
2010	12	12	$928.43

1. Enter the historical data and select the data area (**E13:E24**)
2. Click on **Risk Simulator | Forecasting | Nonlinear Extrapolation**
3. Select the function type and extrapolation periods are required and click **OK**

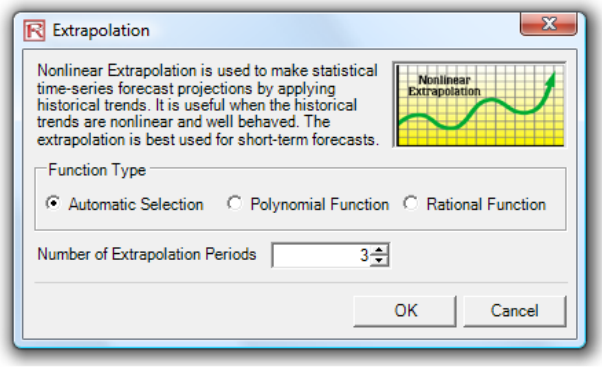


Figure 3.11 – Running a Nonlinear Extrapolation

Nonlinear Extrapolation

Statistical Summary

Extrapolation involves making statistical projections by using historical trends that are projected for a specified period of time into the future. It is only used for time-series forecasts. For cross-sectional or mixed panel data (time-series with cross-sectional data), multivariate regression is more appropriate. This methodology is useful when major changes are not expected, that is, causal factors are expected to remain constant or when the causal factors of a situation are not clearly understood. It also helps discourage introduction of personal biases into the process. Extrapolation is fairly reliable, relatively simple, and inexpensive. However, extrapolation, which assumes that recent and historical trends will continue, produces large forecast errors if discontinuities occur within the projected time period. That is, pure extrapolation of time series assumes that all we need to know is contained in the historical values of the series that is being forecasted. If we assume that past behavior is a good predictor of future behavior, extrapolation is appealing. This makes it a useful approach when all that is needed are many short-term forecasts.

This methodology estimates the f(x) function for any arbitrary x value, by interpolating a smooth nonlinear curve through all the x values, and using this smooth curve, extrapolates future x values beyond the historical data set. The methodology employs either the polynomial functional form or the rational functional form (a ratio of two polynomials). Typically, a polynomial functional form is sufficient for well-behaved data, however, rational functional forms are sometimes more accurate (especially with polar functions, i.e., functions with denominators approaching zero).

Period	Actual	Forecast Fit	Estimate Error
1	1.00		
2	6.73	1.00	
3	20.52	-1.42	-8.15
4	45.25	99.82	119.36
5	83.59	55.92	-46.67
6	138.01	136.71	14.39
7	210.87	211.96	1.69
8	304.44	304.43	-0.41
9	420.89	420.89	0.01
10	562.34	562.34	0.00
11	730.85	730.85	0.00
12	928.43	928.43	0.00
Forecast 13		1157.03	0.00
Forecast 14		1418.57	0.00
Forecast 15		1714.95	0.00
Forecast 16		2048.00	0.00
Forecast 17		2419.55	0.00
Forecast 18		2831.39	0.00

Error Measurements	
RMSE	19.6799
MSE	387.2974
MAD	10.2095
MAPE	31.56%
Theil's U	1.1210

Function Type: Rational

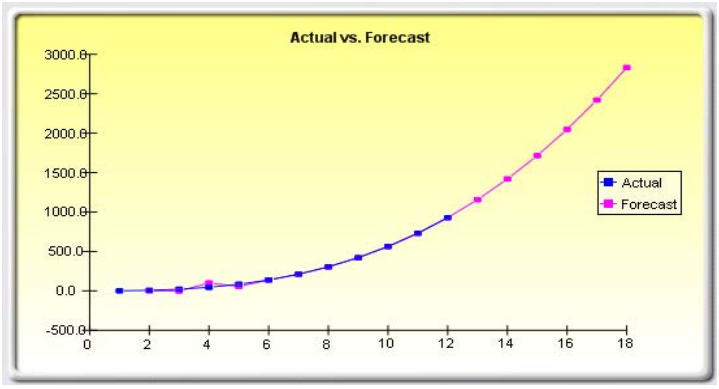


Figure 3.12 – Nonlinear Extrapolation Results

3.7 Box-Jenkins ARIMA Advanced Time-Series

Theory

One very powerful advanced times-series forecasting tool is the ARIMA, or *Auto Regressive Integrated Moving Average,* approach. ARIMA forecasting assembles three separate tools into a comprehensive model. The first tool segment is the autoregressive (AR) term, which corresponds to the number of lagged value of the residual in the unconditional forecast model. In essence, the model captures the historical variation of actual data to a forecasting model and uses this variation or residual to create a better predicting model. The second tool segment is the integration order (I) term. This integration term corresponds to the number of differencing the time series to be forecasted goes through. This element accounts for any nonlinear growth rates existing in the data. The third tool segment is the moving average (MA) term, which is essentially the moving average of lagged forecast errors. By incorporating this lagged forecast errors term, the model in essence learns from its forecast errors or mistakes and corrects for them through a moving-average calculation. The ARIMA model follows the Box-Jenkins methodology with each term representing steps taken in the model construction until only random noise remains. Also, ARIMA modeling uses correlation techniques in generating forecasts. ARIMA can be used to model patterns that may not be visible in plotted data. In addition, ARIMA models can be mixed with exogenous variables, but make sure that the exogenous variables have enough data points to cover the additional number of periods to forecast. Finally, be aware that due to the complexity of the models, this module may take longer to run.

There are many reasons why an ARIMA model is superior to common time-series analysis and multivariate regressions. The common finding in time-series analysis and multivariate regression is that the error residuals are correlated with their own lagged values. This serial correlation violates the standard assumption of regression theory that disturbances are not correlated with other disturbances. The primary problems associated with serial correlation are:

- Regression analysis and basic time-series analysis are no longer efficient among the different linear estimators. However, as the error residuals can help to predict current error residuals, we can take advantage of this information to form a better prediction of the dependent variable using ARIMA.
- Standard errors computed using the regression and time-series formula are not correct, and are generally understated, and if there are lagged-dependent variables set as the regressors, regression estimates are biased and inconsistent but can be fixed using ARIMA.

ARIMA(p,d,q) models are the extension of the AR model that uses three components for modeling the serial correlation in the time series data. The first component is the autoregressive (AR) term. The AR(p) model uses the p lags of the time series in the equation. An AR(p) model has the form: $y_t = a_1 y_{t-1} + ... + a_p y_{t-p} + e_t$. The second component is the integration (d) order term. Each integration order corresponds to differencing the time series. I(1) means differencing the data once; I(d) means differencing the data d times. The third component is the moving average (MA) term. The MA(q) model uses the q lags of the forecast errors to improve the forecast. An MA(q) model has the form: $y_t = e_t + b_1 e_{t-1} + ... + b_q e_{t-q}$. Finally, an ARIMA(p,q) model has the combined form: $y_t = a_1 y_{t-1} + ... + a_p y_{t-p} + e_t + b_1 e_{t-1} + ... + b_q e_{t-q}$.

Procedure

- Start Excel and enter your data or open an existing worksheet with historical data to forecast (the illustration shown next uses the file example file Time-Series ARIMA).
- Select the time-series data and select *Risk Simulator | Forecasting | ARIMA*.
- Enter the relevant *P, D, Q* parameters (positive integers only), enter the number of forecast period desired, and click *OK*.

Notes

For ARIMA and Auto ARIMA, you can model and forecast future periods by either using only the dependent variable (*Y*), that is, the *Time Series Variable* by itself, or you can add in exogenous variables ($X_1, X_2, \ldots, X_n$) just like in a regression analysis where you have multiple independent variables. You can run as many forecast periods as you wish if you use only the time-series variable (*Y*). However, if you add exogenous variables (*X*), note that your forecast period is limited to the number of exogenous variables' data periods minus the time-series variable's data periods. For example, you can only forecast up to 5 periods if you have time-series historical data of 100 periods and only if you have exogenous variables of 105 periods (100 historical periods to match the time-series variable and 5 additional future periods of independent exogenous variables to forecast the time-series dependent variable).

Results Interpretation

In interpreting the results of an ARIMA model, most of the specifications are identical to the multivariate regression analysis (see *Modeling Risk*: *Applying Monte Carlo Simulation, Real Options Analysis, Stochastic Forecasting, and Portfolio Optimization*, *Second Edition*, by Dr. Johnathan Mun for more technical details about interpreting the multivariate regression analysis and ARIMA models). There are however, several additional sets of results specific to the ARIMA analysis as seen in Figure 3.14. The first is the addition of Akaike information criterion (AIC) and Schwarz criterion (SC), which are often used in ARIMA model selection and identification. That is, AIC and SC are used to determine if a particular model with a specific set of p, d, and q parameters is a good statistical fit. SC imposes a greater penalty for additional coefficients than the AIC but, generally, the model with the lowest the AIC and SC values should be chosen. Finally, an additional set of results called the autocorrelation (AC) and partial autocorrelation (PAC) statistics are provided in the ARIMA report.

For instance, if autocorrelation AC(1) is nonzero, it means that the series is first-order serially correlated. If AC dies off more or less geometrically with increasing lags, it implies that the series follows a low-order autoregressive process. If AC drops to zero after a small number of lags, it implies that the series follows a low-order moving-average process. In contrast, PAC measures the correlation of values that are *k* periods apart after removing the correlation from the intervening lags. If the pattern of autocorrelation can be captured by an autoregression of order less than *k*, then the partial autocorrelation at lag *k* will be close to zero. The Ljung-Box Q-statistics and their p-values at lag *k* are also provided, where the null hypothesis being tested is such that there is no autocorrelation up to order *k*. The dotted lines in the plots of the autocorrelations are the approximate two standard error bounds. If the autocorrelation is within these bounds, it is not significantly different from zero at approximately the 5% significance level. Finding the right ARIMA model takes practice and experience. These AC, PAC, SC, and AIC diagnostic tools are highly useful in helping to identify the correct model specification.

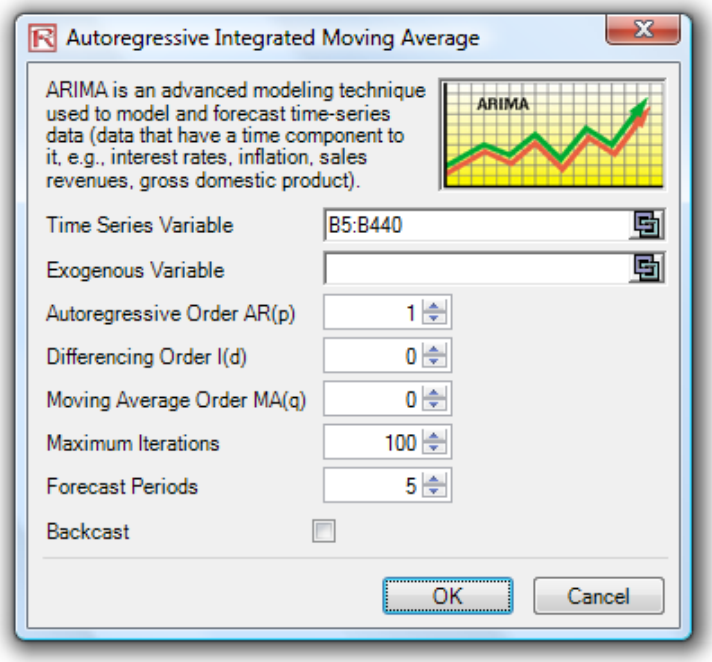


Figure 3.13 – Box-Jenkins ARIMA Forecast Tool

ARIMA (Autoregressive Integrated Moving Average)

Regression Statistics

R-Squared (Coefficient of Determination)	*0.9999*	*Akaike Information Criterion (AIC)*	*4.6213*
Adjusted R-Squared	*0.9999*	*Schwarz Criterion (SC)*	*4.6632*
Multiple R (Multiple Correlation Coefficient)	*1.0000*	*Log Likelihood*	*-1005.1340*
Standard Error of the Estimates (SEy)	*297.5246*	*Durbin-Watson (DW) Statistic*	*1.8588*
Number of Observations	*435*	*Number of Iterations*	*5*

*Autoregressive Integrated Moving Average or ARIMA(p,d,q) models are the extension of the AR model that use three components for modeling the serial correlation in the time-series data. The first component is the autoregressive (AR) term. The AR(p) model uses the p lags of the time series in the equation. An AR(p) model has the form: y(t)=a(1)*y(t-1)+...+a(p)*y(t-p)+e(t).The second component is the integration (d) order term. Each integration order corresponds to differencing the time series. I(1) means differencing the data once. I(d) means differencing the data d times. The third component is the moving average (MA) term. The MA(q) model uses the q lags of the forecast errors to improve the forecast. An MA(q) model has the form: y(t)=e(t)+b(1)*e(t-1)+...+b(q)*e(t-q).Finally, an ARMA(p,q) model has the combined form: y(t)=a(1)*y(t-1)+...+a(p)*y(t-p)+e(t)+b(1)*e(t-1)+...+b(q)*e(t-q).*

The R-Squared, or Coefficient of Determination, indicates the percent variation in the dependent variable that can be explained and accounted for by the independent variables in this regression analysis. However, in a multiple regression, the Adjusted R-Squared takes into account the existence of additional independent variables or regressors and adjusts this R-Squared value to a more accurate view the regression's explanatory power. However, under some ARIMA modeling circumstances (e.g., with nonconvergence models), the R-Squared tends to be unreliable.

The Multiple Correlation Coefficient (Multiple R) measures the correlation between the actual dependent variable (Y) and the estimated or fitted (Y) based on the regression equation. This correlation is also the square root of the Coefficient of Determination (R-Squared).

The Standard Error of the Estimates (SEy) describes the dispersion of data points above and below the regression line or plane. This value is used as part of the calculation to obtain the confidence interval of the estimates later.

The AIC and SC are often used in model selection. SC imposes a greater penalty for additional coefficients. Generally, the user should select a model with the lowest value of the AIC and SC.

The Durbin-Watson statistic measures the serial correlation in the residuals. Generally, DW less than 2 implies positive serial correlation.

Regression Results

	Intercept	*AR(1)*	*MA(1)*
Coefficients	*-0.0626*	*1.0055*	*0.4936*
Standard Error	*0.3108*	*0.0006*	*0.0420*
t-Statistic	*-0.2013*	*1691.1373*	*11.7633*
p-Value	*0.8406*	***0.0000***	***0.0000***
Lower 5%	*0.4498*	*1.0065*	*0.5628*
Upper 95%	*-0.5749*	*1.0046*	*0.4244*

Degrees of Freedom		***Hypothesis Test***	
Degrees of Freedom for Regression	*2*	*Critical t-Statistic (99% confidence with df of 432)*	*2.5873*
Degrees of Freedom for Residual	*432*	*Critical t-Statistic (95% confidence with df of 432)*	*1.9655*
Total Degrees of Freedom	*434*	*Critical t-Statistic (90% confidence with df of 432)*	*1.6484*

The Coefficients provide the estimated regression intercept and slopes. For instance, the coefficients are estimates of the true; population b values in the following regression equation $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$. *The Standard Error measures how accurate the predicted Coefficients are, and the t-Statistics are the ratios of each predicted Coefficient to its Standard Error.*

The t-Statistic is used in hypothesis testing, where we set the null hypothesis (Ho) such that the real mean of the Coefficient = 0, and the alternate hypothesis (Ha) such that the real mean of the Coefficient is not equal to 0. A t-test is is performed and the calculated t-Statistic is compared to the critical values at the relevant Degrees of Freedom for Residual. The t-test is very important as it calculates if each of the coefficients is statistically significant in the presence of the other regressors. This means that the t-test statistically verifies whether a regressor or independent variable should remain in the regression or it should be dropped.

The Coefficient is statistically significant if its calculated t-Statistic exceeds the Critical t-Statistic at the relevant degrees of freedom (df). The three main confidence levels used to test for significance are 90%, 95% and 99%. If a Coefficient's t-Statistic exceeds the Critical level, it is considered statistically significant. Alternatively, the p-Value calculates each t-Statistic's probability of occurrence, which means that the smaller the p-Value, the more significant the Coefficient. The usual significant levels for the p-Value are 0.01, 0.05, and 0.10, corresponding to the 99%, 95%, and 99% confidence levels.

The Coefficients with their p-Values highlighted in blue indicate that they are statistically significant at the 90% confidence or 0.10 alpha level, while those highlighted in red indicate that they are not statistically significant at any other alpha levels.

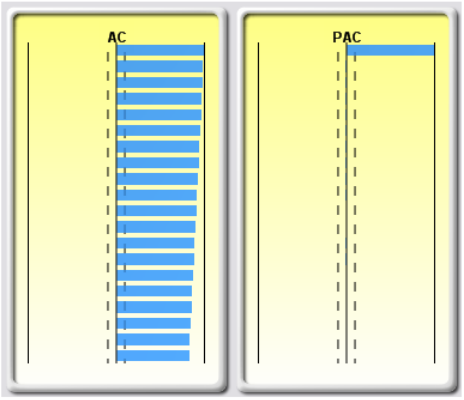
Analysis of Variance

	Sums of Squares	**Mean of Squares**	**F-Statistic**	**p-Value**	***Hypothesis Test***	
Regression	*38415447.5277*	*19207723.7638*	*3171851.1034*	***0.0000***	*Critical F-statistic (99% confidence with df of 2 and 432)*	*4.6546*
Residual	*2616.0549*	*6.0557*			*Critical F-statistic (95% confidence with df of 2 and 432)*	*3.0166*
Total	*38418063.5826*	*19207729.8195*			*Critical F-statistic (90% confidence with df of 2 and 432)*	*2.3149*

The Analysis of Variance (ANOVA) table provides an F-test of the regression model's overall statistical significance. Instead of looking at individual regressors as in the t-test, the F-test looks at all the estimated Coefficients' statistical properties. The F-Statistic is calculated as the ratio of the Regression's Mean of Squares to the Residual's Mean of Squares. The numerator measures how much of the regression is explained, while the denominator measures how much is unexplained. Hence, the larger the F-Statistic, the more significant the model. The corresponding p-Value is calculated to test the null hypothesis (Ho) where all the Coefficients are simultaneously equal to zero, versus the alternate hypothesis (Ha) that they are all simultaneously different from zero, indicating a significant overall regression model. If the p-Value is smaller than the 0.01, 0.05, or 0.10 alpha significance, then the regression is significant. The same approach can be applied to the F-Statistic by comparing the calculated F-Statistic with the critical F values at various significance levels.

Autocorrelation

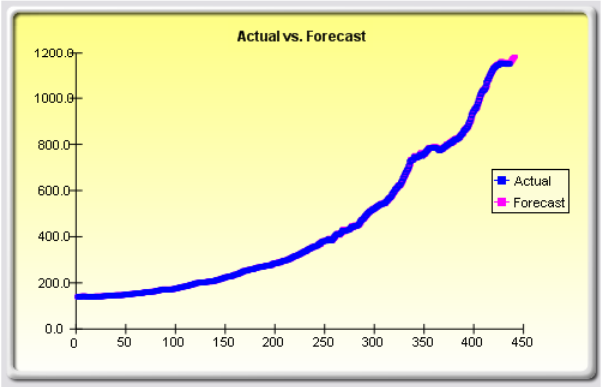
Time Lag	AC	PAC	Lower Bound	Upper Bound	Q-Stat	Prob
1	0.9921	0.9921	(0.0958)	0.0958	431.1216	-
2	0.9841	(0.0105)	(0.0958)	0.0958	856.3037	-
3	0.9760	(0.0109)	(0.0958)	0.0958	1,275.4818	-
4	0.9678	(0.0142)	(0.0958)	0.0958	1,688.5499	-
5	0.9594	(0.0098)	(0.0958)	0.0958	2,095.4625	-
6	0.9509	(0.0113)	(0.0958)	0.0958	2,496.1572	-
7	0.9423	(0.0124)	(0.0958)	0.0958	2,890.5594	-
8	0.9336	(0.0147)	(0.0958)	0.0958	3,278.5669	-
9	0.9247	(0.0121)	(0.0958)	0.0958	3,660.1152	-
10	0.9156	(0.0139)	(0.0958)	0.0958	4,035.1192	-
11	0.9066	(0.0049)	(0.0958)	0.0958	4,403.6117	-
12	0.8975	(0.0068)	(0.0958)	0.0958	4,765.6032	-
13	0.8883	(0.0097)	(0.0958)	0.0958	5,121.0697	-
14	0.8791	(0.0087)	(0.0958)	0.0958	5,470.0032	-
15	0.8698	(0.0064)	(0.0958)	0.0958	5,812.4256	-
16	0.8605	(0.0056)	(0.0958)	0.0958	6,148.3694	-
17	0.8512	(0.0062)	(0.0958)	0.0958	6,477.8620	-
18	0.8419	(0.0038)	(0.0958)	0.0958	6,800.9622	-
19	0.8326	(0.0003)	(0.0958)	0.0958	7,117.7709	-
20	0.8235	0.0002	(0.0958)	0.0958	7,428.3952	-



If autocorrelation AC(1) is nonzero, it means that the series is first order serially correlated. If AC(k) dies off more or less geometrically with increasing lag , it implies that the series follows a low-order autoregressive process. If AC(k) drops to zero after a small number of lags, it implies that the series follows a low-order moving-average process. Partial correlation PAC(k) measures the correlation of values that are k periods apart after removing the correlation from the intervening lags. If the pattern of autocorrelation can be captured by an autoregression of order less than k, then the partial autocorrelation at lag k will be close to zero. Ljung-Box Q-statistics and their p-values at lag k has the null hypothesis that there is no autocorrelation up to order k. The dotted lines in the plots of the autocorrelations are the approximate two standard error bounds. If the autocorrelation is within these bounds, it is not significantly different from zero at (approximately) the 5% significance level.

Forecasting

Period	Actual (Y)	Forecast (F)	Error (E)
2	139.399994	139.6056	(0.2056)
3	139.699997	140.0069	(0.3069)
4	139.699997	140.2586	(0.5586)
5	140.699997	140.1343	0.5657
6	141.199997	141.6948	(0.4948)
7	141.699997	141.6741	0.0259
8	141.899994	142.4339	(0.5339)
9	141	142.3587	(1.3587)
10	140.5	141.0466	(0.5466)
11	140.399994	140.9447	(0.5447)
12	140	140.8451	(0.8451)
13	140	140.2946	(0.2946)
14	139.899994	140.5663	(0.6663)
15	139.800003	140.2823	(0.4823)
16	139.600006	140.2726	(0.6726)
17	139.600006	139.9775	(0.3775)
18	139.600006	140.1232	(0.5231)
19	140.199997	140.0513	0.1487
20	141.300003	140.9862	0.3138
21	141.199997	142.1738	(0.9738)
22	140.899994	141.4377	(0.5377)
23	140.899994	141.3513	(0.4513)
24	140.699997	141.3939	(0.6939)
25	141.100006	141.0731	0.0270
26	141.600006	141.8311	(0.2311)
27	141.899994	142.2065	(0.3065)
28	142.100006	142.4709	(0.3709)
29	142.699997	142.6402	0.0598
30	142.899994	143.4561	(0.5561)
31	142.899994	143.3532	(0.4532)
32	143.5	143.4040	0.0960
33	143.800003	144.2784	(0.4784)
34	144.100006	144.2966	(0.1966)
35	144.800003	144.7374	0.0626
36	145.199997	145.5692	(0.3692)
37	145.199997	145.7582	(0.5582)
38	145.699997	145.6649	0.0351
39	146	146.4605	(0.4605)
40	146.399994	146.5176	(0.1176)
41	146.800003	147.0891	(0.2891)
42	146.600006	147.4066	(0.8066)



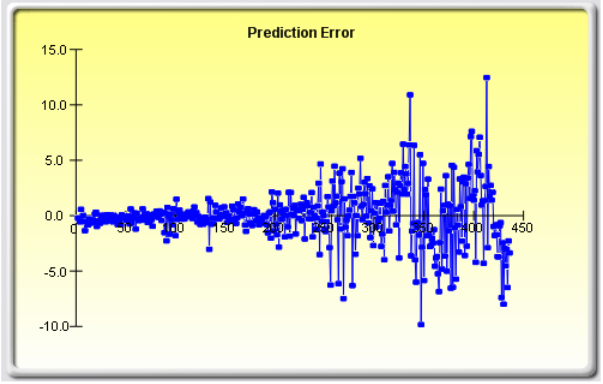


Figure 3.14 – Box-Jenkins ARIMA Forecast Report

3.8 AUTO-ARIMA (Box-Jenkins ARIMA Advanced Time-Series)

Theory

While the analyses are identical, AUTO ARIMA differs from ARIMA in automating some of the traditional ARIMA modeling. It automatically tests multiple permutations of model specifications and returns the best-fitting model. Running the Auto ARIMA is similar to regular ARIMA forecasting, with the difference being that the P, D, Q inputs are no longer required and different combinations of these inputs are automatically run and compared.

Procedure

- Start Excel and enter your data or open an existing worksheet with historical data to forecast (the illustration shown in Figure 3.15 uses the example file Advanced Forecasting Models in the Examples menu of Risk Simulator).
- In the Auto ARIMA worksheet, select the data and click on *Risk Simulator | Forecasting | AUTO-ARIMA*. You can also access this method through the forecasting icons ribbon, or right-clicking anywhere in the model and selecting the forecasting shortcut menu.
- Click on the link icon and link to the existing time-series data, enter the number of forecast periods desired, and click *OK*.

Notes

For ARIMA and Auto ARIMA, you can model and forecast future periods by either using only the dependent variable (Y), that is, the *Time Series Variable* by itself or you can add in exogenous variables ($X_1, X_2, \ldots, X_n$) just like in a regression analysis where you have multiple independent variables. You can run as many forecast periods as you wish if you use only the time-series variable (Y). However, if you add exogenous variables (X), note that your forecast period is limited to the number of exogenous variables' data periods minus the time-series variable's data periods. For example, you can only forecast up to 5 periods if you have time-series historical data of 100 periods and only if you have exogenous variables of 105 periods (100 historical periods to match the time-series variable and 5 additional future periods of independent exogenous variables to forecast the time-series dependent variable).

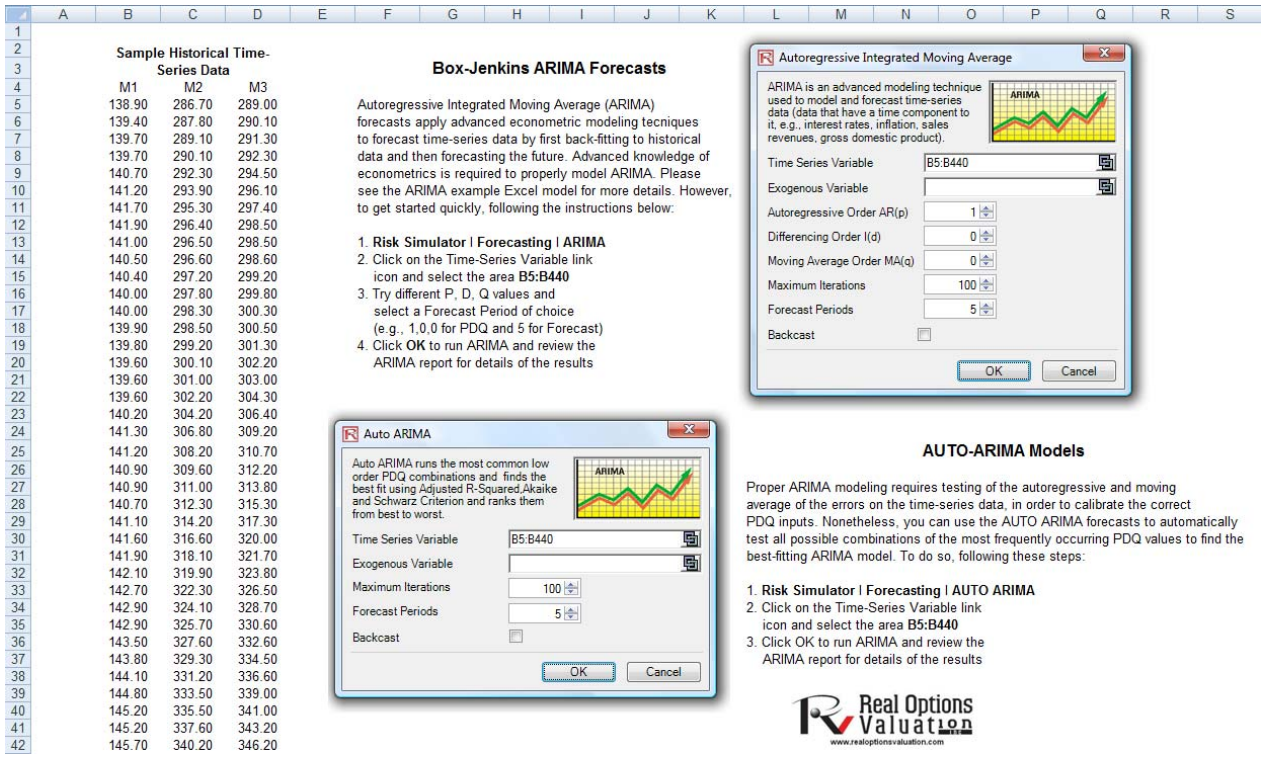


Figure 3.15 – AUTO ARIMA Module

3.9 Basic Econometrics

Theory

Econometrics refers to a branch of business analytics, modeling, and forecasting techniques for modeling the behavior or forecasting certain business or economic variables. Running the *Basic Econometrics* models is similar to regular regression analysis except that the dependent and independent variables are allowed to be modified before a regression is run. The report generated and its interpretation is the same as shown in the Multivariate Regression section presented earlier.

Procedure

- Start Excel and enter your data or open an existing worksheet with historical data to forecast (the illustration shown in Figure 3.16 uses the file example file Advanced Forecasting Models in the Examples menu of Risk Simulator).
- Select the data in the *Basic Econometrics* worksheet and select *Risk Simulator | Forecasting | Basic Econometrics*.
- Enter the desired dependent and independent variables (see Figure 3.16 for examples) and click *OK* to run the model and report, or click on *Show Results* to view the results before generating the report in case you need to make any changes to the model

Basic Econometrics Data Set

Y	X1	X2	X3	X4	X5
521	18308	185	4.041	79.6	7.2
367	1148	600	0.55	1	8.5
443	18068	372	3.665	32.3	5.7
365	7729	142	2.351	45.1	7.3
614	100484	432	29.76	190.8	7.5
385	16728	290	3.294	31.8	5
286	14630	346	3.287	678.4	6.7
397	4008	328	0.666	340.8	6.2
764	38927	354	12.938	239.6	7.3
427	22322	266	6.478	111.9	5
153	3711	320	1.108	172.5	2.8
231	3136	197	1.007	12.2	6.1
524	50508	266	11.431	205.6	7.1
328	28886	173	5.544	154.6	5.9
240	16996	190	2.777	49.7	4.6
286	13035	239	2.478	30.3	4.4
285	12973	190	3.685	92.8	7.4
569	16309	241	4.22	96.9	7.1
96	5227	189	1.228	39.8	7.5
498	19235	358	4.781	489.2	5.9
481	44487	315	6.016	767.6	9
468	44213	303	9.295	163.6	9.2
177	23619	228	4.375	55	5.1
198	9106	134	2.573	54.9	8.6
458	24917	189	5.117	74.3	6.6
108	3872	196	0.799	5.5	6.9
246	8945	183	1.578	20.5	2.7
291	2373	417	1.202	10.9	5.5
68	7128	233	1.109	123.7	7.2
311	23624	349	7.73	1042	6.6
606	5242	284	1.515	12.5	6.9
512	92629	499	17.99	381	7.2
426	28795	231	6.629	136.1	5.8
47	4487	143	0.639	9.3	4.1
265	48799	249	10.847	264.9	6.4
370	14067	195	3.146	45.8	6.7
312	12693	288	2.842	29.6	6
222	62184	229	11.882	265.1	6.9
280	9153	287	1.003	960.3	8.5

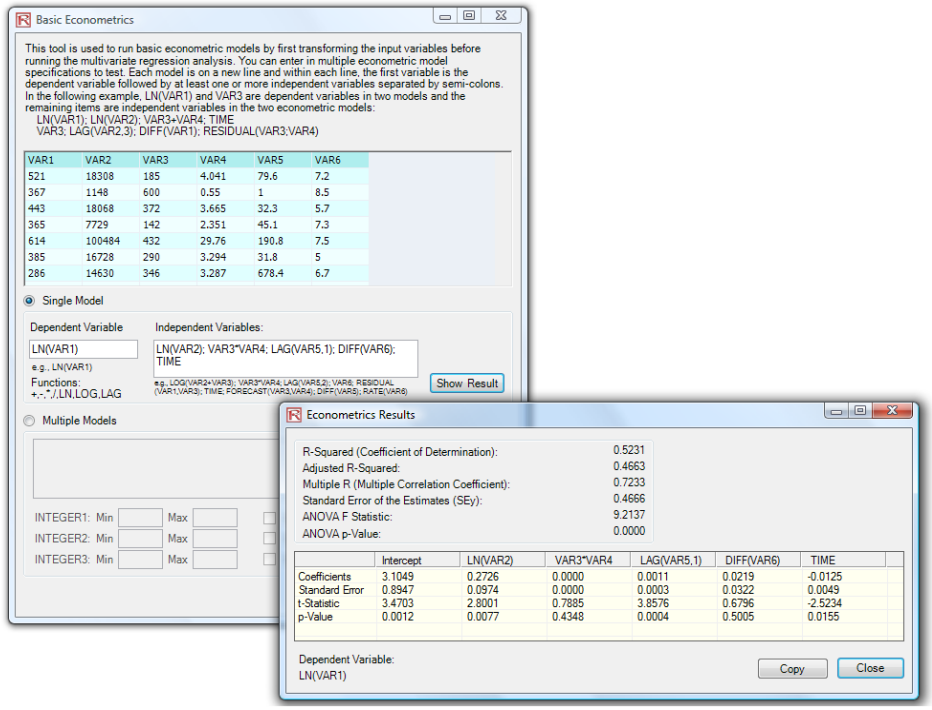


Figure 3.16 – Basic Econometrics Module

Notes

- To run an econometric model, simply select the data (B5:G55) including headers and click on *Risk Simulator | Forecasting | Basic Econometrics*. You can then type in the variables and their modifications for the dependent and independent variables (Figure 3.16). Note that only one variable is allowed as the *Dependent Variable (Y)*, whereas multiple variables are allowed in the *Independent Variables (X)* section, separated by a semicolon (;), and that basic mathematical functions can be used (e.g., LN, LOG, LAG, +, -, /, *, TIME, RESIDUAL, DIFF). Click on *Show Results* to preview the computed model and click *OK* to generate the econometric model report.

- You can also automatically generate *Multiple Models* by entering a sample model and using the predefined *INTEGER(N)* variable as well as *Shifting Data* up or down specific rows repeatedly. For instance, if you use the variable *LAG(VAR1, INTEGER1)* and you set *INTEGER1* to be between *MIN* = 1 and *MAX* = 3, then the following three models will be run: *LAG(VAR1,1)*, then *LAG(VAR1,2)*, and, finally, *LAG(VAR1,3)*. Also, sometimes you might want to test if the time-series data has structural shifts or if the behavior of the model is consistent over time by shifting the data and then running the same model. For example, if you have 100 months of data listed chronologically, you can shift it down 3 months at a time for 10 times (i.e., the model will be run on months 1–100, 4–100, 7–100, etc.). Using this *Multiple Models* section in *Basic Econometrics*, you can run hundreds of models by simply entering a single model equation if you use these predefined integer variables and shifting methods.

3.10 J-S Curve Forecasts

Theory

The J curve, or exponential growth curve, is one where the growth of the next period depends on the current period's level and the increase is exponential. This means that over time, the values will increase significantly, from one period to another. This model is typically used in forecasting biological growth and chemical reactions over time.

Procedure

- Start Excel and select *Risk Simulator | Forecasting | JS Curves*.
- Select the J or S curve type, *enter the required input assumptions* (see Figures 3.17 and 3.18 for examples), and click *OK* to run the model and report.

The S curve, or logistic growth curve, starts off like a J curve, with exponential growth rates. Over time, the environment becomes saturated (e.g., market saturation, competition, overcrowding), the growth slows, and the forecast value eventually ends up at a saturation or maximum level. This model is typically used in forecasting market share or sales growth of a new product from market introduction until maturity and decline, population dynamics, growth of bacterial cultures, and other naturally occurring variables. Figure 3.18 illustrates a sample S curve.

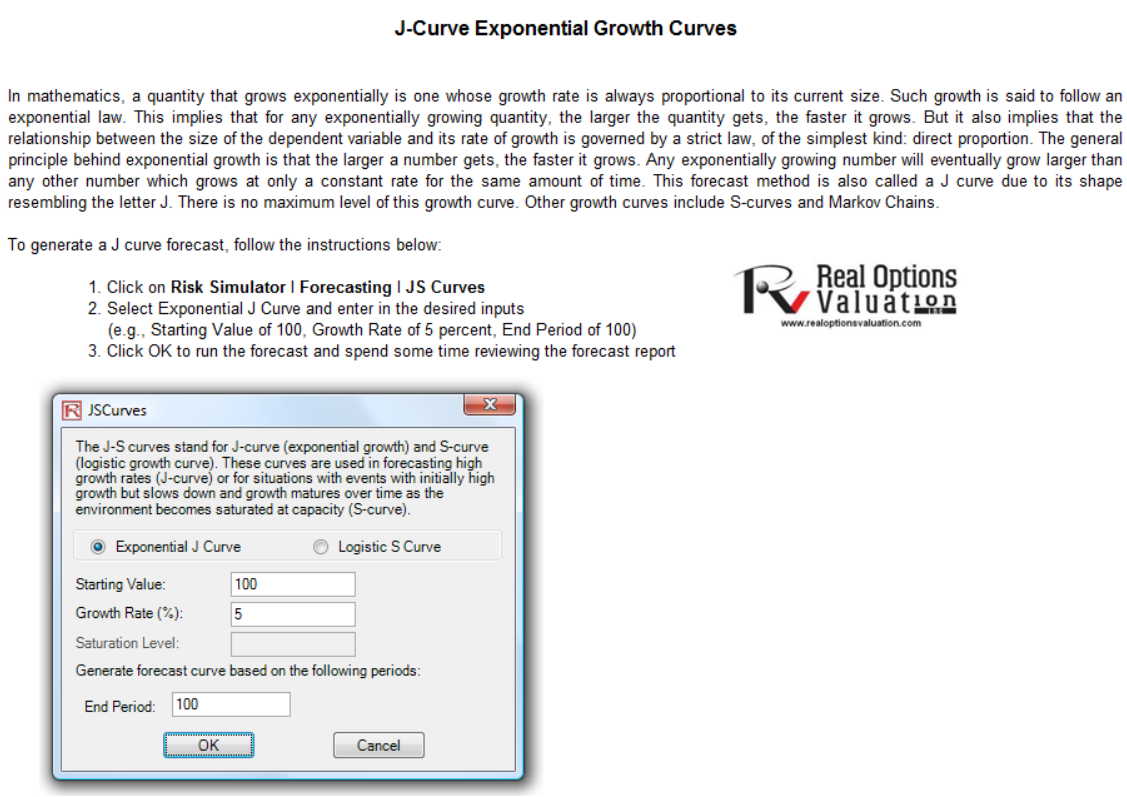


Figure 3.17 – J-Curve Forecast

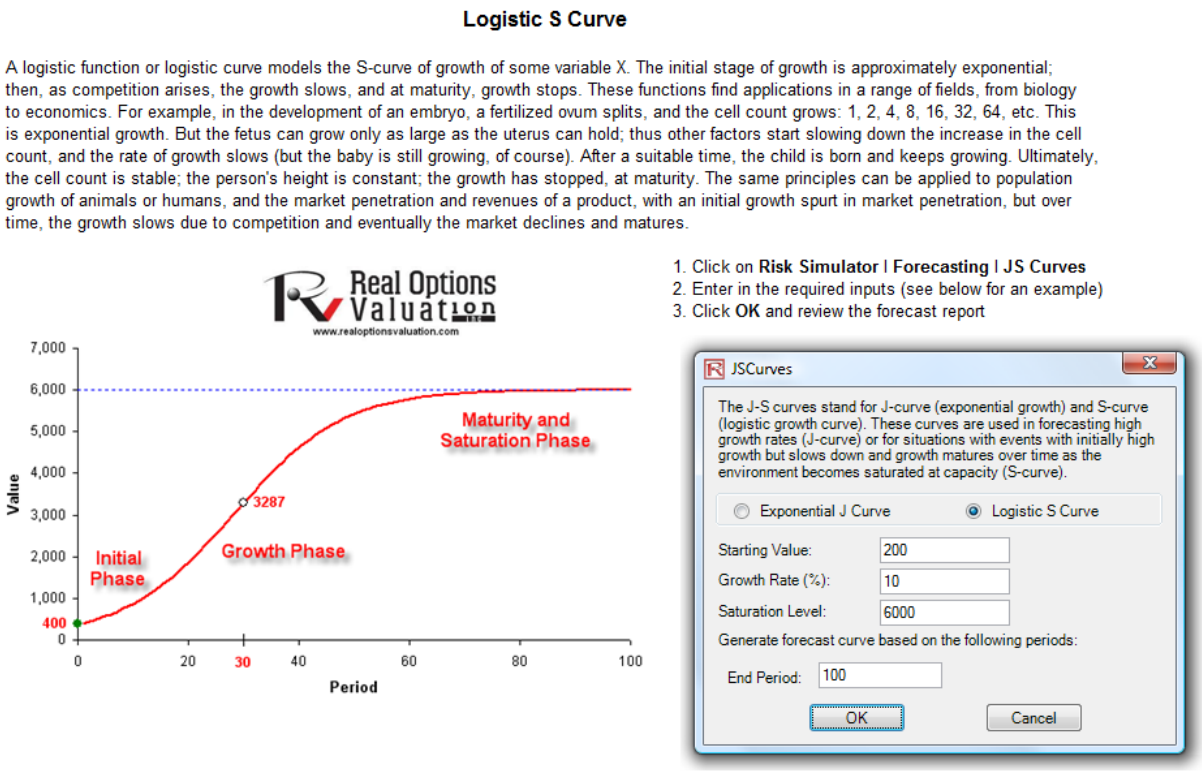


Figure 3.18 – S-Curve Forecast

3.11 GARCH Volatility Forecasts

Theory

The generalized autoregressive conditional heteroskedasticity (GARCH) model is used to model historical and forecast future volatility levels of a marketable security (e.g., stock prices, commodity prices, oil prices, etc.). The data set has to be a time series of raw price levels. GARCH will first convert the prices into relative returns and then run an internal optimization to fit the historical data to a mean-reverting volatility term structure, while assuming that the volatility is heteroskedastic in nature (changes over time according to some econometric characteristics). The theoretical specifics of a GARCH model are outside the purview of this user manual. For more details on GARCH models, please refer to *Advanced Analytical Models*, by Dr. Johnathan Mun (Wiley Finance, 2008).

Procedure

- Start Excel and open the example file *Advanced Forecasting Model,* go to the *GARCH* worksheet, select the data and click on *Risk Simulator | Forecasting | GARCH*.
- Click on the link icon, select the *Data Location*, enter the required input assumptions (see Figure 3.19), and click *OK* to run the model and report.

Note: The typical volatility forecast situation requires P = 1, Q = 1, Periodicity = number of periods per year (12 for monthly data, 52 for weekly data, 252 or 365 for daily data), Base = minimum of 1 and up to the periodicity value, and Forecast Periods = number of annualized volatility forecasts you wish to obtain. There are several GARCH models available in Risk Simulator, including EGARCH, EGARCH-T, GARCH-M, GJR-GARCH, GJR-GARCH-T, IGARCH, and T-GARCH. See the chapter in *Modeling Risk, Second Edition*, by Dr. Johnathan Mun (Wiley Finance, 2010), on GARCH modeling for more details on what each specification is for.

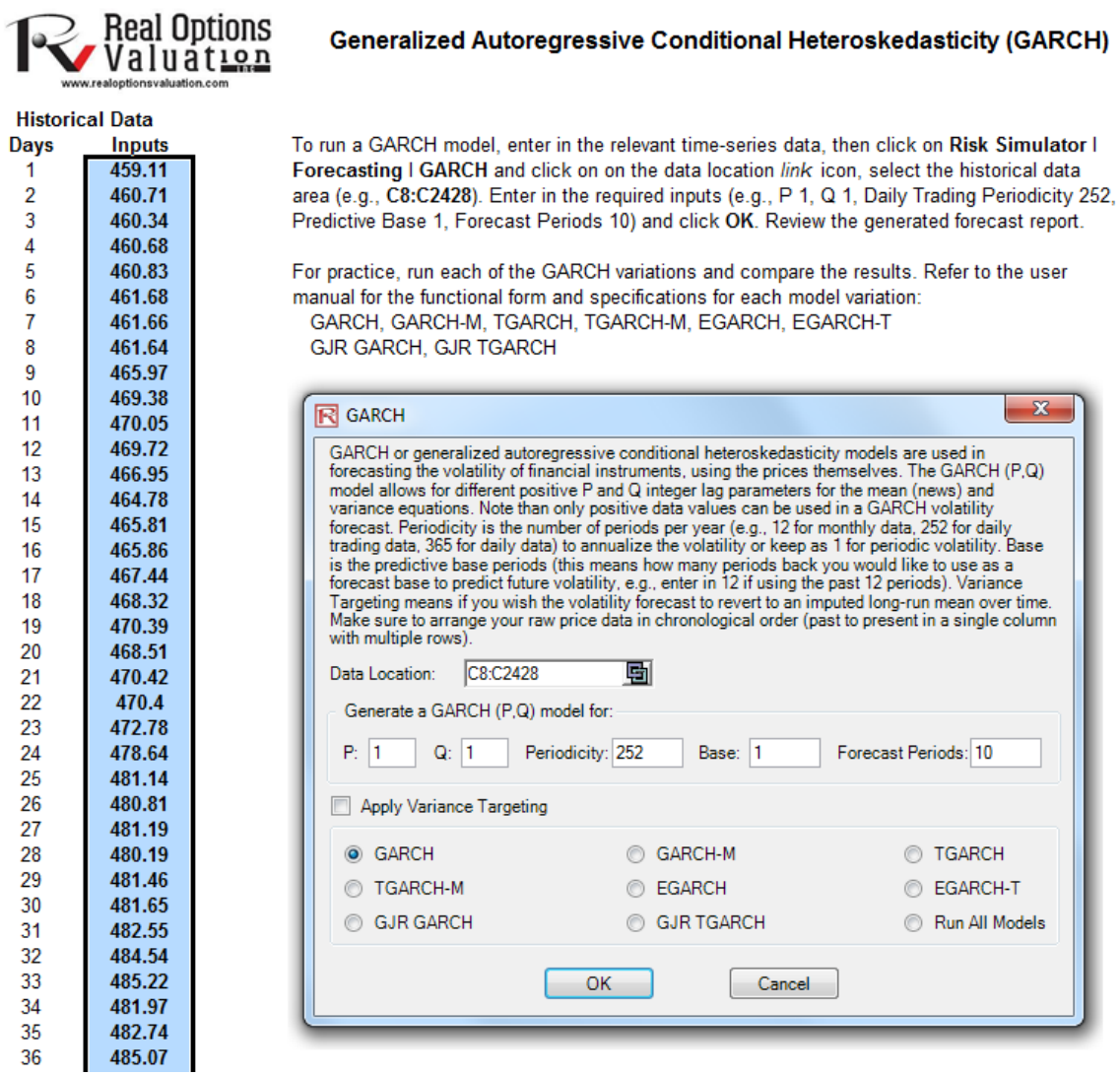


Figure 3.19 – GARCH Volatility Forecast

3.11.1 GARCH Equations

The accompanying table lists some of the GARCH specifications used in Risk Simulator with two underlying distributional assumptions: one for normal distribution and the other for the t distribution.

	z_t ~ **Normal Distribution**	z_t ~ T-Distribution
GARCH-M Variance in Mean Equation	$y_t = c + \lambda\sigma_t^2 + \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$	$y_t = c + \lambda\sigma_t^2 + \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$
GARCH-M Standard Deviation in Mean Equation	$y_t = c + \lambda\sigma_t + \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$	$y_t = c + \lambda\sigma_t + \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$
GARCH-M Log Variance in Mean Equation	$y_t = c + \lambda\ln(\sigma_t^2) + \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$	$y_t = c + \lambda\ln(\sigma_t^2) + \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$
GARCH	$y_t = x_t\gamma + \varepsilon_t$ $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$	$y_t = \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$
EGARCH	$y_t = \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\ln\left(\sigma_t^2\right) = \omega + \beta\cdot\ln\left(\sigma_{t-1}^2\right) +$ $\alpha\left[\left\lvert\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right\rvert - E(\lvert\varepsilon_t\rvert)\right] + r\frac{\varepsilon_{t-1}}{\sigma_{t-1}}$ $E(\lvert\varepsilon_t\rvert) = \sqrt{\frac{2}{\pi}}$	$y_t = \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\ln\left(\sigma_t^2\right) = \omega + \beta\cdot\ln\left(\sigma_{t-1}^2\right) +$ $\alpha\left[\left\lvert\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right\rvert - E(\lvert\varepsilon_t\rvert)\right] + r\frac{\varepsilon_{t-1}}{\sigma_{t-1}}$ $E(\lvert\varepsilon_t\rvert) = \frac{2\sqrt{\nu-2}\,\Gamma((\nu+1)/2)}{(\nu-1)\Gamma(\nu/2)\sqrt{\pi}}$
GJR-GARCH	$y_t = \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 +$ $r\varepsilon_{t-1}^2 d_{t-1} + \beta\sigma_{t-1}^2$ $d_{t-1} = \begin{cases} 1 & \text{if } \varepsilon_{t-1} < 0 \\ 0 & \text{otherwise} \end{cases}$	$y_t = \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 +$ $r\varepsilon_{t-1}^2 d_{t-1} + \beta\sigma_{t-1}^2$ $d_{t-1} = \begin{cases} 1 & \text{if } \varepsilon_{t-1} < 0 \\ 0 & \text{otherwise} \end{cases}$

For the GARCH-M models, the conditional variance equations are the same in the six variations but the mean questions are different and assumption on z_t can be either normal distribution or t distribution. The estimated parameters for GARCH-M with normal distribution are those five parameters in the mean and conditional variance equations. The estimated parameters for GARCH-M with the t distribution are those five parameters in the mean and conditional variance equations plus another parameter, the degrees of freedom for the t distribution. In contrast, for the GJR models, the mean equations are the same in the six variations and the differences are that the conditional variance equations and the assumption on z_t can be either a normal distribution or t distribution. The estimated parameters for EGARCH and GJR-GARCH with normal distribution are those four parameters in the conditional variance equation. The estimated parameters for GARCH, EARCH, and GJR-GARCH with t distribution are those parameters in the conditional variance equation plus the degrees of freedom for the t distribution. More technical details of GARCH methodologies fall outside of the scope of this book.

3.12 Markov Chains

Theory

A Markov chain exists when the probability of a future state depends on a previous state and when linked together form a chain that reverts to a long-run steady state level. This approach is typically used to forecast the market share of two competitors. The required inputs are the starting probability of a customer in the first store (the first state) will return to the same store in the next period versus the probability of switching to a competitor's store in the next state.

Procedure

- Start Excel and select *Risk Simulator | Forecasting | Markov Chain*.
- *Enter in the required input assumptions* (see Figure 3.20 for an example) and click *OK* to run the model and report.

Notes

Set both probabilities to 10% and rerun the Markov chain and you will see the effects of switching behaviors very clearly in the resulting chart.

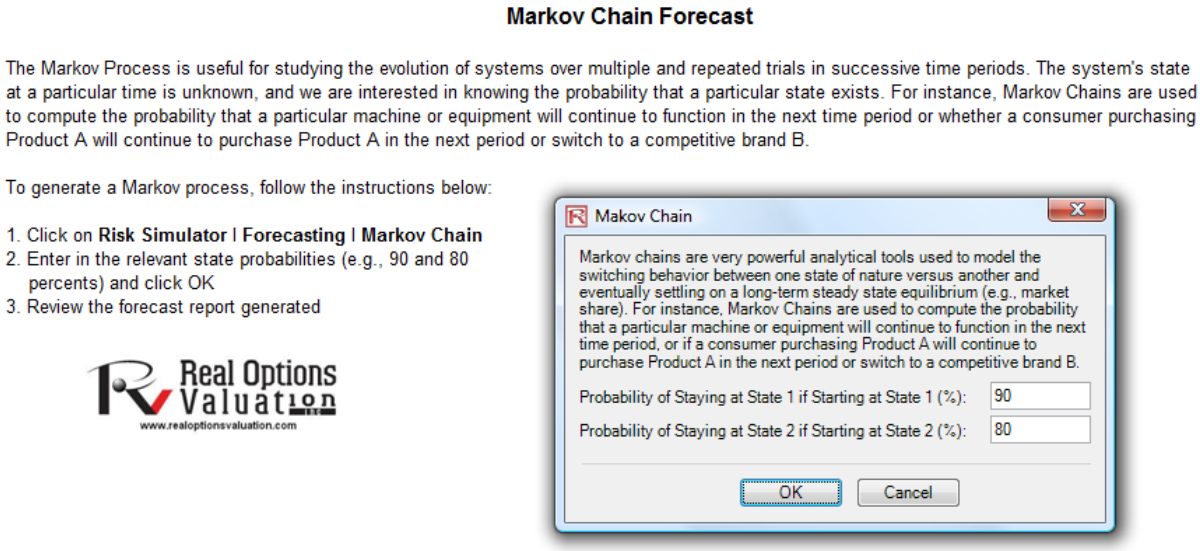


Figure 3.20 – Markov Chains (Switching Regimes)

3.13 Limited Dependent Variables: Logit, Probit, Tobit (Maximum Likelihood Estimation)

Theory

The term *Limited Dependent Variables* describes the situation where the dependent variable contains data that are limited in scope and range, such as binary responses (*0* or *1*) or truncated, ordered, or censored data. For instance, given a set of independent variables (e.g., age, income, education level of credit card or mortgage loan holders), we can model the probability of default using maximum likelihood estimation (MLE). The response, or dependent variable *Y,* is binary. That is, it can have only two possible outcomes that we denote as *1* and *0* (e.g., *Y* may represent presence/absence of a certain condition, defaulted/not defaulted on previous loans, success/failure of some device, answer yes/no on a survey, etc.). We also have a vector of independent variable regressors *X,* which are assumed to influence the outcome *Y*. A typical ordinary least squares regression approach is invalid because the regression errors are heteroskedastic and non-normal, and the resulting estimated probability estimates will return nonsensical values of above *1* or below *0*. MLE analysis handles these problems using an iterative optimization routine to maximize a log likelihood function when the dependent variables are limited.

A Logit or Logistic regression, is used for predicting the probability of occurrence of an event by fitting data to a logistic curve. It is a generalized linear model used for binomial regression, and, like many forms of regression analysis, it makes use of several predictor variables that may be either numerical or categorical. MLE applied in a binary multivariate logistic analysis is used to model dependent variables to determine the expected probability of success of belonging to a certain group. The estimated coefficients for the Logit model are the logarithmic odds ratios and cannot be interpreted directly as probabilities. A quick computation is first required and the approach is simple.

Specifically, the Logit model is specified as $Estimated\ Y = LN[P_i/(1-P_i)]$ or, conversely, $P_i = EXP(Estimated\ Y)/(1+EXP(Estimated\ Y))$, and the coefficients β_i are the log odds ratios. So,

taking the antilog, or $EXP(\beta_i)$, we obtain the odds ratio of $P_i/(1-P_i)$. This means that with an increase in a unit of βi, the log odds ratio increases by this amount. Finally, the rate of change is the probability $dP/dX = \beta_i P_i(1-P_i)$. The standard error measures how accurate the predicted coefficients are, and the t-statistics are the ratios of each predicted coefficient to its standard error and are used in the typical regression hypothesis test of the significance of each estimated parameter. To estimate the probability of success of belonging to a certain group (e.g., predicting if a smoker will develop chest complications given the amount smoked per year), simply compute the *Estimated Y* value using the MLE coefficients. For example, if the model is $Y = 1.1 + 0.005$ *(Cigarettes),* then someone smoking 100 packs per year has an *Estimated Y* of $1.1 + 0.005(100) = 1.6$. Next, compute the inverse antilog of the odds ratio by $EXP(Estimated\ Y)/[1 + EXP(Estimated\ Y)] = EXP(1.6)/(1+ EXP(1.6)) = 0.8320$. So, such a person has an *83.20%* chance of developing some chest complications in his or her lifetime.

A Probit model (sometimes also known as a Normit model) is a popular alternative specification for a binary response model, which employs a probit function estimated using maximum likelihood estimation and the approach is called probit regression. The Probit and Logistic regression models tend to produce very similar predictions where the parameter estimates in a logistic regression tend to be 1.6 to 1.8 times higher than they are in a corresponding Probit model. The choice of using a Probit or Logit is entirely up to convenience, and the main distinction is that the logistic distribution has a higher kurtosis (fatter tails) to account for extreme values. For example, suppose that house ownership is the decision to be modeled, and this response variable is binary (home purchase or no home purchase) and depends on a series of independent variables X_i such as income, age, and so forth, such that $I_i = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n$, where the larger the value of I_i, the higher the probability of home ownership. For each family, a critical I^* threshold exists where, if exceeded, the house is purchased, otherwise, no home is purchased, and the outcome probability (P) is assumed to be normally distributed such that $P_i = CDF(I)$ using a standard normal cumulative distribution function (*CDF*). Therefore, using the estimated coefficients exactly like those of a regression model and using the *Estimated Y* value, apply a standard normal distribution (you can use Excel's *NORMSDIST* function or Risk Simulator's *Distributional Analysis* tool by selecting *Normal* distribution and setting the mean to be *0* and standard deviation to be *1*). Finally, to obtain a Probit or probability unit measure, set $I_i + 5$ (because whenever the probability $P_i < 0.5$, the estimated I_i is negative, due to the fact that the normal distribution is symmetrical around a mean of zero).

The Tobit Model (Censored Tobit) is an econometric and biometric modeling method used to describe the relationship between a non-negative dependent variable Y_i and one or more independent variables X_i. The dependent variable in a Tobit econometric model is censored; it is censored because values below zero are not observed. The Tobit model assumes that there is a latent unobservable variable Y^*. This variable is linearly dependent on the X_i variables via a vector of β_i coefficients that determine their interrelationships. In addition, there is a normally distributed error term U_i to capture random influences on this relationship. The observable variable Y_i is defined to be equal to the latent variables whenever the latent variables are above zero and is assumed to be zero otherwise. That is, $Y_i = Y^*$ if $Y^* > 0$ and $Y_i = 0$ if $Y^* = 0$. If the relationship parameter β_i is estimated by using ordinary least squares regression of the observed Y_i on X_i, the resulting regression estimators are inconsistent and yield downward-biased slope coefficients and an upward-biased intercept. Only MLE would be consistent for a Tobit model. In the Tobit model, there is an ancillary statistic called *sigma*, which is equivalent to the standard error of estimate in a standard ordinary least squares regression, and the estimated coefficients are used the same way as a regression analysis.

Procedure

- Start Excel and open the example file *Advanced Forecasting Model*, go to the *MLE* worksheet, *select the data* set including the headers, and click on *Risk Simulator | Forecasting | Maximum Likelihood*.
- *Select the dependent variable* from the drop-down list (see Figure 3.21) and click *OK* to run the model and report.

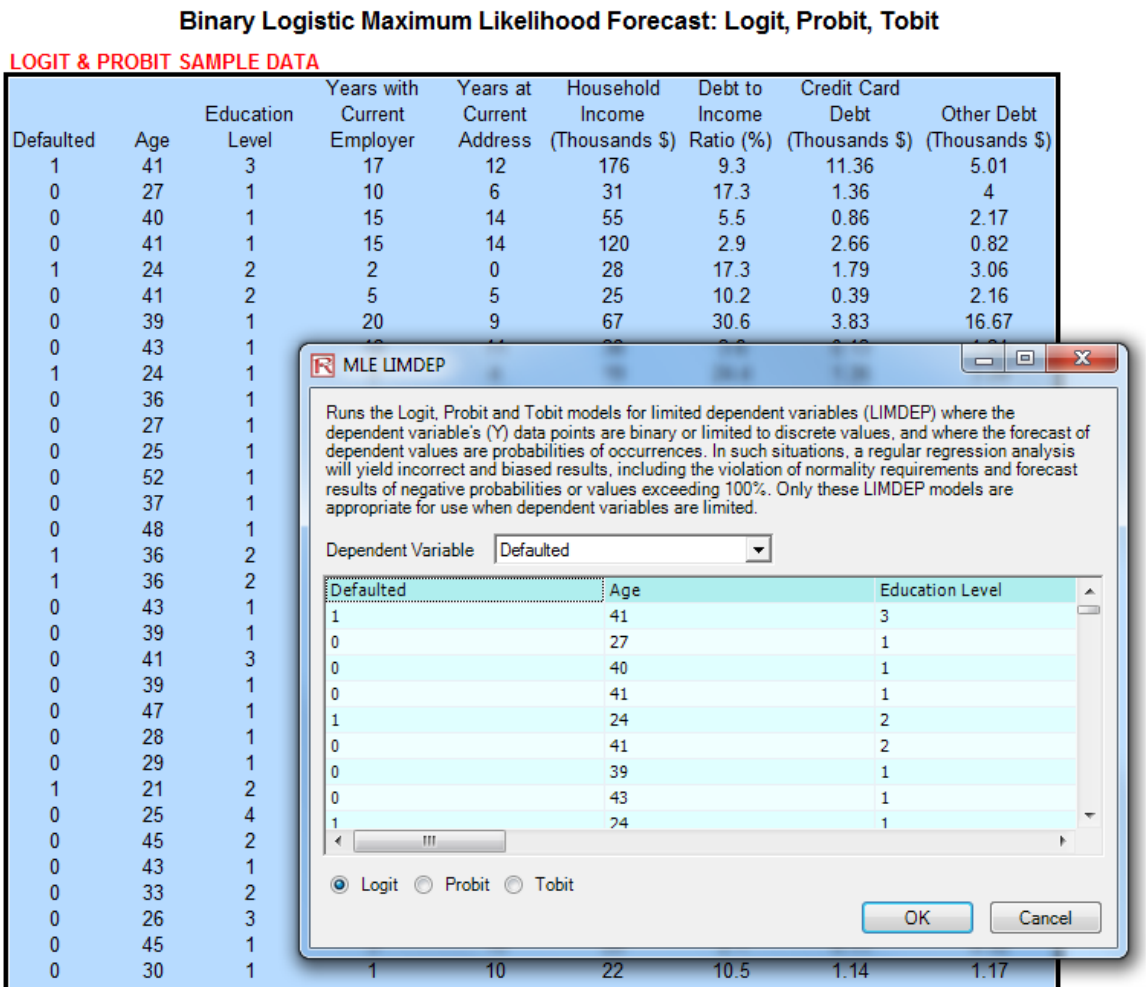
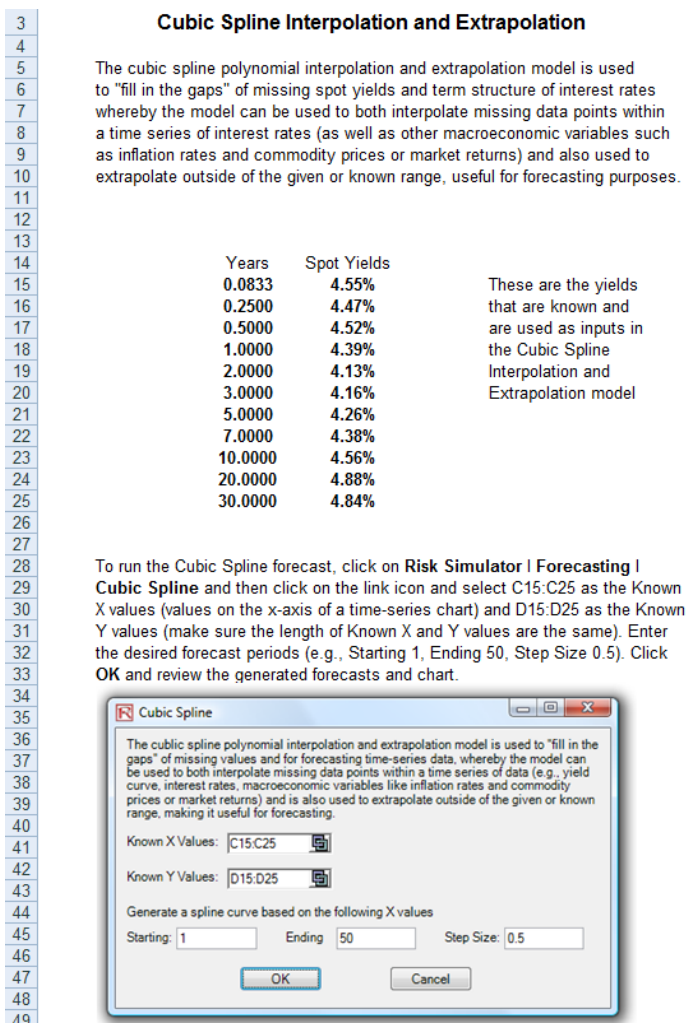


Figure 3.21 – Maximum Likelihood Module

3.14 Spline (Cubic Spline Interpolation and Extrapolation)

Theory

Sometimes there are missing values in a time-series data set. For instance, interest rates for years 1 to 3 may exist, followed by years 5 to 8, and then year 10. Spline curves can be used to interpolate the missing years' interest rate values based on the data that exist. Spline curves can also be used to forecast or extrapolate values of future time periods beyond the time period of available data. The data can be linear or nonlinear. Figure 3.22 illustrates how a cubic spline is run and Figure 3.23 shows the resulting forecast report from this module. The Known *X* values represent the values on the *x*-axis of a chart (in our example, this is Years of the known interest rates, and, usually, the *x*-axis values are those that are known in advance such as time or years) and the Known *Y* values represent the values on the *y*-axis (in our case, the known Interest Rates). The *y*-axis variable is typically the variable you wish to interpolate missing values from or extrapolate the values into the future.



Cubic Spline Interpolation and Extrapolation

The cubic spline polynomial interpolation and extrapolation model is used to "fill in the gaps" of missing spot yields and term structure of interest rates whereby the model can be used to both interpolate missing data points within a time series of interest rates (as well as other macroeconomic variables such as inflation rates and commodity prices or market returns) and also used to extrapolate outside of the given or known range, useful for forecasting purposes.

Years	Spot Yields
0.0833	4.55%
0.2500	4.47%
0.5000	4.52%
1.0000	4.39%
2.0000	4.13%
3.0000	4.16%
5.0000	4.26%
7.0000	4.38%
10.0000	4.56%
20.0000	4.88%
30.0000	4.84%

These are the yields that are known and are used as inputs in the Cubic Spline Interpolation and Extrapolation model

To run the Cubic Spline forecast, click on **Risk Simulator** | **Forecasting** | **Cubic Spline** and then click on the link icon and select C15:C25 as the Known X values (values on the x-axis of a time-series chart) and D15:D25 as the Known Y values (make sure the length of Known X and Y values are the same). Enter the desired forecast periods (e.g., Starting 1, Ending 50, Step Size 0.5). Click **OK** and review the generated forecasts and chart.

Figure 3.22 – Cubic Spline Module

Procedure

- Start Excel and open the example file *Advanced Forecasting Model,* go to the *Cubic Spline* worksheet, *select the data* set excluding the headers, and click on *Risk Simulator | Forecasting | Cubic Spline.*
- The data location is automatically inserted into the user interface if you first select the data, or you can also manually click on the link icon and link the *Known X values* and *Known Y values* (see Figure 3.22 for an example), then enter in the required *Starting* and *Ending* values to extrapolate and interpolate, as well as the required *Step Size* between these starting and ending values. Click *OK* to run the model and report (see Figure 3.23).

Cubic Spline Forecasts

The cubic spline polynomial interpolation and extrapolation model is used to "fill in the gaps" of missing values and for forecasting time-series data, whereby the model can be used to both interpolate missing data points within a time series of data (e.g., yield curve, interest rates, macroeconomic variables like inflation rates and commodity prices or market returns) and also used to extrapolate outside of the given or known range, making it useful for forecasting.

Spline Interpolation and Extrapolation Results

X	Fitted Y	Notes
1.0	4.39%	Interpolate
1.5	4.21%	Interpolate
2.0	4.13%	Interpolate
2.5	4.13%	Interpolate
3.0	4.16%	Interpolate
3.5	4.19%	Interpolate
4.0	4.22%	Interpolate
4.5	4.24%	Interpolate
5.0	4.26%	Interpolate
5.5	4.29%	Interpolate
6.0	4.32%	Interpolate
6.5	4.35%	Interpolate
7.0	4.38%	Interpolate
7.5	4.41%	Interpolate
8.0	4.44%	Interpolate
8.5	4.47%	Interpolate
9.0	4.50%	Interpolate
9.5	4.53%	Interpolate
10.0	4.56%	Interpolate
10.5	4.59%	Interpolate

Real Options Valuation
www.realoptionsvaluation.com

These are the known value inputs in the Cublic Spline Interpolation and Extrapolation model:

Observation	Known X	Known Y
1	0.0833	4.55%
2	0.2500	4.47%
3	0.5000	4.52%
4	1.0000	4.39%
5	2.0000	4.13%
6	3.0000	4.16%
7	5.0000	4.26%
8	7.0000	4.38%
9	10.0000	4.56%
10	20.0000	4.88%
11	30.0000	4.84%

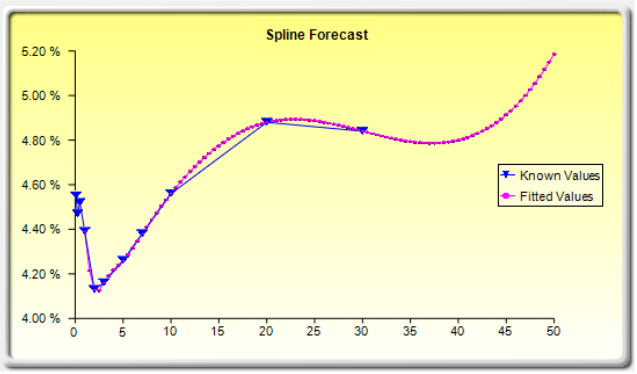


Figure 3.23 – Spline Forecast Results

4

4. OPTIMIZATION

This chapter looks at the optimization process and methodologies in more detail in connection with using Risk Simulator. These methodologies include the use of continuous versus discrete integer optimization, as well as static versus dynamic and stochastic optimizations.

4.1 Optimization Methodologies

Many algorithms exist to run optimization, and many different procedures exist when optimization is coupled with Monte Carlo simulation. In Risk Simulator, there are three distinct optimization procedures and optimization types as well as different decision variable types. For instance, Risk Simulator can handle Continuous Decision Variables (1.2535, 0.2215, etc.) as well as Integers Decision Variables (1, 2, 3, 4, etc.), Binary Decision Variables (1 and 0 for go and no-go decisions), and Mixed Decision Variables (both integers and continuous variables). On top of that, Risk Simulator can handle Linear Optimization (i.e., when both the objective and constraints are all linear equations and functions) as well as Nonlinear Optimizations (i.e., when the objective and constraints are a mixture of linear and nonlinear functions and equations).

Static Optimization

As far as the optimization process is concerned, Risk Simulator can be used to run a Static Optimization, that is, an optimization that is run on a static model, where no simulations are run. In other words, all the inputs in the model are static and unchanging. This optimization type is applicable when the model is assumed to be known and no uncertainties exist. Also, a discrete optimization can be first run to determine the optimal portfolio and its corresponding optimal allocation of decision variables before more advanced optimization procedures are applied. For instance, before running a stochastic optimization problem, a discrete optimization is first run to determine if there exist solutions to the optimization problem before a more protracted analysis is performed.

Dynamic Optimization

Next, Dynamic Optimization is applied when Monte Carlo simulation is used together with optimization. Another name for such a procedure is Simulation-Optimization. That is, a simulation is first run, then the results of the simulation are then applied in the Excel model, and then an optimization is applied to the simulated values. In other words, a simulation is run for N trials, and then an optimization process is run for M iterations until the optimal results are obtained or an infeasible set is found. That is, using Risk Simulator's optimization module, you can choose which forecast and assumption statistics to use and replace in the model after the simulation is run. Then, these forecast statistics can be applied in the optimization process. This approach is useful when you have a large model with many interacting assumptions and forecasts, and when some of the forecast statistics are required in the optimization. For example, if the standard deviation of an assumption or forecast is required in the optimization

model (e.g., computing the Sharpe ratio in asset allocation and optimization problems where we have mean divided by standard deviation of the portfolio), then this approach should be used.

Stochastic Optimization

The Stochastic Optimization process, in contrast, is similar to the dynamic optimization procedure with the exception that the entire dynamic optimization process is repeated T times. That is, a simulation with N trials is run, and then an optimization is run with M iterations to obtain the optimal results. Then the process is replicated T times. The results will be a forecast chart of each decision variable with T values. In other words, a simulation is run and the forecast or assumption statistics are used in the optimization model to find the optimal allocation of decision variables. Then, another simulation is run, generating different forecast statistics, and these new updated values are then optimized, and so forth. Hence, the final decision variables will each have their own forecast chart, indicating the range of the optimal decision variables. For instance, instead of obtaining single-point estimates in the dynamic optimization procedure, you can now obtain a distribution of the decision variables and, hence, a range of optimal values for each decision variable, also known as a stochastic optimization.

Efficient Frontier

Finally, an Efficient Frontier optimization procedure applies the concepts of marginal increments and shadow pricing in optimization. That is, what would happen to the results of the optimization if one of the constraints were relaxed slightly? Say, for instance, the budget constraint is set at $1 million. What would happen to the portfolio's outcome and optimal decisions if the constraint were now $1.5 million, or $2 million, and so forth? This is the concept of the Markowitz efficient frontiers in investment finance, whereby one can determine what additional returns the portfolio will generate if the portfolio standard deviation is allowed to increase slightly. This process is similar to the dynamic optimization process with the exception that one of the constraints is allowed to change, and with each change, the simulation and optimization process is run. This process is best applied manually using Risk Simulator. That is, run a dynamic or stochastic optimization, then rerun another optimization with a constraint, and repeat that procedure several times. This manual process is important because by changing the constraint, the analyst can determine if the results are similar or different, and, hence, whether it is worthy of any additional analysis, or the analyst can determine how far a marginal increase in the constraint should be to obtain a significant change in the objective and decision variables.

One item is worthy of consideration. There exist other software products that supposedly perform stochastic optimization but, in fact, they do not. For instance, after a simulation is run, then one iteration of the optimization process is generated, and then another simulation is run, then the second optimization iteration is generated and so forth. This approach is simply a waste of time and resources. That is, in optimization, the model is put through a rigorous set of algorithms, where multiple iterations (ranging from several to thousands of iterations) are required to obtain the optimal results. Hence, generating one iteration at a time is a waste of time and resources. The same portfolio can be solved using Risk Simulator in under a minute as compared to multiple hours using such a backward approach. Also, such a simulation-optimization approach will typically yield bad results, and it is not a stochastic optimization approach. Be extremely careful of such methodologies when applying optimization to your models.

The next two sections provide examples of optimization problems. One uses continuous decision variables while the other uses discrete integer decision variables. In either model, you can apply discrete optimization, dynamic optimization, stochastic optimization, or even the efficient frontiers with shadow pricing. Any of these approaches can be used for these two examples. Therefore, for simplicity, only the model setup is illustrated and it is up to the user to decide which optimization process to run. Also, the continuous model uses the nonlinear

optimization approach (because the portfolio risk computed is a nonlinear function, and the objective is a nonlinear function of portfolio returns divided by portfolio risks) and integer optimization is an example of a linear optimization model (its objective and all of its constraints are linear). Therefore, these two examples encapsulate all of the procedures aforementioned.

4.2 Optimization with Continuous Decision Variables

Figure 4.1 illustrates the sample continuous optimization model. The example here uses the Continuous Optimization file found either on the start menu at Start | Real Options Valuation | Risk Simulator | Examples or accessed directly through Risk Simulator | Example Models. In this example, there are 10 distinct asset classes (e.g., different types of mutual funds, stocks, or assets) where the idea is to most efficiently and effectively allocate the portfolio holdings such that the best bang for the buck is obtained; that is, to generate the best portfolio returns possible given the risks inherent in each asset class. To truly understand the concept of optimization, we will have to delve deeply into this sample model to see how the optimization process can best be applied.

As mentioned, the model shows the 10 asset classes each with its own set of annualized returns and annualized volatilities. These return and risk measures are annualized values such that they can be consistently compared across different asset classes. Returns are computed using the geometric average of the relative returns, while the risks are computed using the logarithmic relative stock returns approach.

ASSET ALLOCATION OPTIMIZATION MODEL

Asset Class Description	Annualized Returns	Volatility Risk	Allocation Weights	Required Minimum Allocation	Required Maximum Allocation	Return to Risk Ratio	Returns Ranking (Hi-Lo)	Risk Ranking (Lo-Hi)	Return to Risk Ranking (Hi-Lo)	Allocation Ranking (Hi-Lo)
Asset Class 1	10.54%	12.36%	10.00%	5.00%	35.00%	0.8524	9	2	7	1
Asset Class 2	11.25%	16.23%	10.00%	5.00%	35.00%	0.6929	7	8	10	1
Asset Class 3	11.84%	15.64%	10.00%	5.00%	35.00%	0.7570	6	7	9	1
Asset Class 4	10.64%	12.35%	10.00%	5.00%	35.00%	0.8615	8	1	5	1
Asset Class 5	13.25%	13.28%	10.00%	5.00%	35.00%	0.9977	5	4	2	1
Asset Class 6	14.21%	14.39%	10.00%	5.00%	35.00%	0.9875	3	6	3	1
Asset Class 7	15.53%	14.25%	10.00%	5.00%	35.00%	1.0898	1	5	1	1
Asset Class 8	14.95%	16.44%	10.00%	5.00%	35.00%	0.9094	2	9	4	1
Asset Class 9	14.16%	16.50%	10.00%	5.00%	35.00%	0.8584	4	10	6	1
Asset Class 10	10.06%	12.50%	10.00%	5.00%	35.00%	0.8045	10	3	8	1
Portfolio Total	*12.6419%*	*4.58%*	*100.00%*							
Return to Risk Ratio	*2.7596*									

Specifications of the optimization model:

Objective: *Maximize Return to Risk Ratio (C18)*
Decision Variables: *Allocation Weights (E6:E15)*
Restrictions on Decision Variables: *Minimum and Maximum Required (F6:G15)*
Constraints: *Portfolio Total Allocation Weights 100% (E17 is set to 100%)*

Additional specifications:

1. One can always maximize portfolio total returns or minimize the portfolio total risk.
2. Incorporate Monte Carlo simulation in the model by simulating the returns and volatility of each asset class and apply Simulation-Optimization techniques.
3. The portfolio can be optimized as is without simulation using Static Optimization techniques.

Figure 4.1 – Continuous Optimization Model

Referring to Figure 4.1, column E (Allocation Weights) holds the decision variables, which are the variables that need to be tweaked and tested such that the total weight is constrained at 100% (cell E17). Typically, to start the optimization, we set these cells to a uniform value, where in this case, cells E6 to E15 are set at 10% each. In addition, each decision variable may have specific restrictions in its allowed range. In this example, the lower and upper allocations allowed are 5% and 35%, as seen in columns F and G. This means that each asset class may have its own allocation boundaries. Next, column H shows the return to risk ratio, which is simply the return percentage divided by the risk percentage, where the higher this value, the higher the bang for the buck. Columns I through L show the individual asset class rankings by returns, risk, return to risk ratio, and allocation. In other words, these rankings show at a glance which asset class has the lowest risk, or the highest return, and so forth.

The portfolio's total returns in cell C17 is SUMPRODUCT(C6:C15, E6:E15), that is, the sum of the allocation weights multiplied by the annualized returns for each asset class. In other words, we have $R_P = \omega_A R_A + \omega_B R_B + \omega_C R_C + \omega_D R_D$, where R_P is the return on the portfolio, $R_{A,B,C,D}$ are the individual returns on the projects, and $\omega_{A,B,C,D}$ are the respective weights, or capital allocation, across each project.

In addition, the portfolio's diversified risk in cell D17 is computed by taking: $\sigma_P = \sqrt{\sum_{i=1}^{i} \omega_i^2 \sigma_i^2 + \sum_{i=1}^{n} \sum_{j=1}^{m} 2\omega_i \omega_j \rho_{i,j} \sigma_i \sigma_j}$.

Here, $\rho_{i,j}$ are the respective cross-correlations between the asset classes—hence, if the cross-correlations are negative, there are risk diversification effects, and the portfolio risk decreases. However, to simplify the computations here, we assume zero correlations among the asset classes through this portfolio risk computation, but assume the correlations when applying simulation on the returns as will be seen later. Therefore, instead of applying static correlations among these different asset returns, we apply the correlations in the simulation assumptions themselves, creating a more dynamic relationship among the simulated return values.

Finally, the return to risk ratio, or Sharpe ratio, is computed for the portfolio. This value is seen in cell C18, and represents the objective to be maximized in this optimization exercise. To summarize, we have the following specifications in this example model:

Objective:	*Maximize Return to Risk Ratio (C18)*
Decision Variables:	*Allocation Weights (E6:E15)*
Restrictions on Decision Variables:	*Minimum and Maximum Required (F6:G15)*
Constraints:	*Total Allocation Weights Sum to 100% (E17)*

Procedure

Open the example file and start a new profile by clicking on Risk Simulator | New Profile and provide it a name.

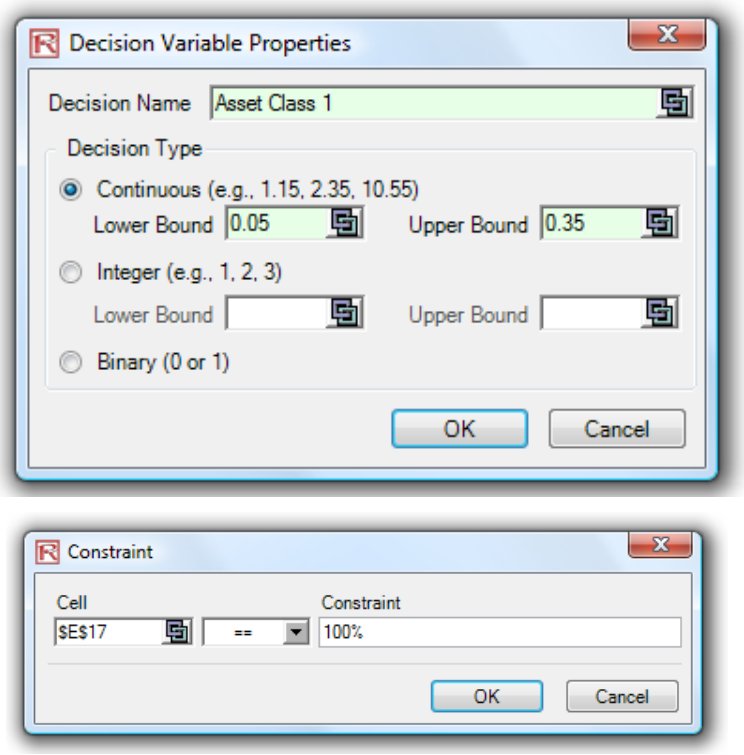
- The first step in optimization is to set the decision variables. Select cell *E6*, set the first decision variable (*Risk Simulator | Optimization | Set Decision*), and click on the link icon to select the name cell (B6), as well as the lower bound and upper bound values at cells *F6* and *G6*. Then, using *Risk Simulator's copy*, copy this cell *E6* decision variable and paste it to the remaining cells in *E7 to E15*.

- The second step in optimization is to set the constraint. There is only one constraint here, that is, the total allocation in the portfolio must sum to 100%. So, click on *Risk Simulator | Optimization | Constraints…* and select *ADD* to add a new constraint. Then, select the cell *E17* and make it *equal (=)* to *100%*. Click *OK* when done.

- The final step in optimization is to set the objective function and start the optimization by selecting the objective cell *C18* and *Risk Simulator | Optimization | Run Optimization* and then selecting the optimization of choice (*Static Optimization, Dynamic Optimization*, or *Stochastic Optimization*). To get started, select *Static Optimization*. Check to make sure the objective cell is set for *C18* and select *Maximize*. You can now review the decision variables and constraints if required, or click *OK* to run the static optimization.

Once the optimization is complete, you may select Revert to revert back to the original values of the decision variables as well as the objective, or select Replace to apply the optimized decision variables. Typically, Replace is chosen after the optimization is done.

Figure 4.2 shows the screen shots of these procedural steps. You can add simulation assumptions on the model's returns and risk (columns C and D) and apply the dynamic optimization and stochastic optimization for additional practice.



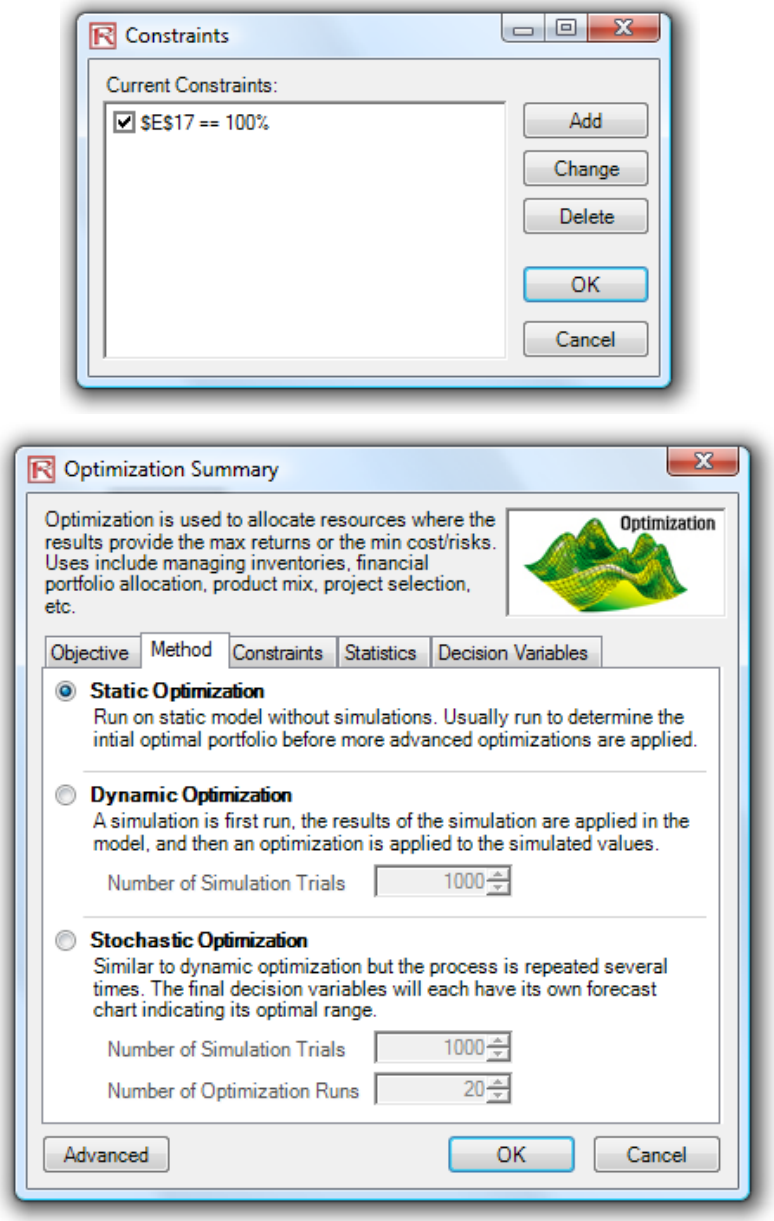


Figure 4.2 – Running Continuous Optimization in Risk Simulator

Results Interpretation

The optimization's final results are shown in Figure 4.3, where the optimal allocation of assets for the portfolio is seen in cells E6:E15. That is, given the restrictions of each asset fluctuating between 5% and 35%, and where the sum of the allocation must equal 100%, the allocation that maximizes the return to risk ratio can be identified from the data provided in Figure 4.3.

A few important things have to be noted when reviewing the results and optimization procedures performed thus far:

- The correct way to run the optimization is to maximize the bang for the buck, or returns to risk Sharpe ratio, as we have done.

- If instead we maximized the total portfolio returns, the optimal allocation result is trivial and does not require optimization to obtain. That is, simply allocate 5% (the minimum allowed) to the lowest eight assets, 35% (the maximum allowed) to the highest returning asset, and the remaining (25%) to the second-best returns asset. Optimization is not required. However, when allocating the portfolio this way, the risk is a lot higher as compared to when maximizing the returns to risk ratio, although the portfolio returns by themselves are higher.
- In contrast, one can minimize the total portfolio risk, but the returns will now be less.

Table 4.1 illustrates the results from the three different objectives being optimized and shows that the best approach is to maximize the returns to risk ratio, that is, for the same amount of risk, this allocation provides the highest amount of return. Conversely, for the same amount of return, this allocation provides the lowest amount of risk possible. This approach of bang for the buck, or returns to risk ratio, is the cornerstone of the Markowitz efficient frontier in modern portfolio theory. That is, if we constrained the total portfolio risk level and successively increased it over time, we will obtain several efficient portfolio allocations for different risk characteristics. Thus, different efficient portfolio allocations can be obtained for different individuals with different risk preferences.

Objective	Portfolio Returns	Portfolio Risk	Portfolio Returns to Risk Ratio
Maximize Returns to Risk Ratio	12.69%	4.52%	2.8091
Maximize Returns	13.97%	6.77%	2.0636
Minimize Risk	12.38%	4.46%	2.7754

Table 4.1 – Optimization Results

ASSET ALLOCATION OPTIMIZATION MODEL

Asset Class Description	Annualized Returns	Volatility Risk	Allocation Weights	Required Minimum Allocation	Required Maximum Allocation	Return to Risk Ratio	Returns Ranking (Hi-Lo)	Risk Ranking (Lo-Hi)	Return to Risk Ranking (Hi-Lo)	Allocation Ranking (Hi-Lo)
Asset Class 1	10.54%	12.36%	11.09%	5.00%	35.00%	0.8524	9	2	7	4
Asset Class 2	11.25%	16.23%	6.87%	5.00%	35.00%	0.6929	7	8	10	10
Asset Class 3	11.84%	15.64%	7.78%	5.00%	35.00%	0.7570	6	7	9	9
Asset Class 4	10.64%	12.35%	11.22%	5.00%	35.00%	0.8615	8	1	5	3
Asset Class 5	13.25%	13.28%	12.08%	5.00%	35.00%	0.9977	5	4	2	2
Asset Class 6	14.21%	14.39%	11.04%	5.00%	35.00%	0.9875	3	6	3	5
Asset Class 7	15.53%	14.25%	12.30%	5.00%	35.00%	1.0898	1	5	1	1
Asset Class 8	14.95%	16.44%	8.90%	5.00%	35.00%	0.9094	2	9	4	7
Asset Class 9	14.16%	16.50%	8.37%	5.00%	35.00%	0.8584	4	10	6	8
Asset Class 10	10.06%	12.50%	10.35%	5.00%	35.00%	0.8045	10	3	8	6
Portfolio Total	*12.6920%*	*4.52%*	*100.00%*							
Return to Risk Ratio	*2.8091*									

Figure 4.3 – Continuous Optimization Results

4.3 Optimization with Discrete Integer Variables

Sometimes, the decision variables are not continuous but are discrete integers (e.g., 0 and 1). We can use optimization with discrete integer variables as on-off switches or go/no-go decisions. Figure 4.4 illustrates a project selection model with 12 projects listed. The example here uses the Discrete Optimization file found either on the start menu at *Start | Real Options Valuation | Risk Simulator | Examples* or accessed directly through Risk Simulator | Example Models. Each project has its own returns (ENPV and NPV, for expanded net present value and net present value—the ENPV is simply the NPV plus any strategic real options values), costs of implementation, risks, and so forth. If required, this model can be modified to include required full-time equivalences (FTE) and other resources of various functions, and additional constraints can be set on these additional resources. The inputs into this model are typically linked from other spreadsheet models. For instance, each project will have its own discounted cash flow or returns on investment model. The application here is to maximize the portfolio's Sharpe ratio subject to some budget allocation. Many other versions of this model can be created, for instance, maximizing the portfolio returns or minimizing the risks, or adding constraints where the total number of projects chosen cannot exceed 6, and so forth and so on. All of these items can be run using this existing model.

Procedure

Open the example file and start a new profile by clicking on *Risk Simulator | New Profile* and provide it a name.

- The first step in optimization is to set up the decision variables. Set the first decision variable by selecting cell *J4*, select *Risk Simulator | Optimization | Set Decision*, click on the link icon to select the name cell (*B4*), and select the *Binary* variable. Then, using *Risk Simulator's copy*, copy this cell *J4* decision variable and paste the decision variable to the remaining cells in *J5 to J15*. This is the best method if you have only several decision variables and you can name each decision variable with a unique name for identification later.

- The second step in optimization is to set the constraint. There are two constraints here: the total budget allocation in the portfolio must be less than $5,000 and the total number of projects must not exceed 6. So, click on *Risk Simulator | Optimization | Constraints…* and select *ADD* to add a new constraint. Then, select the cell *D17* and make it less than or equal to *(<=) 5000*. Repeat by setting cell *J17 <= 6*.

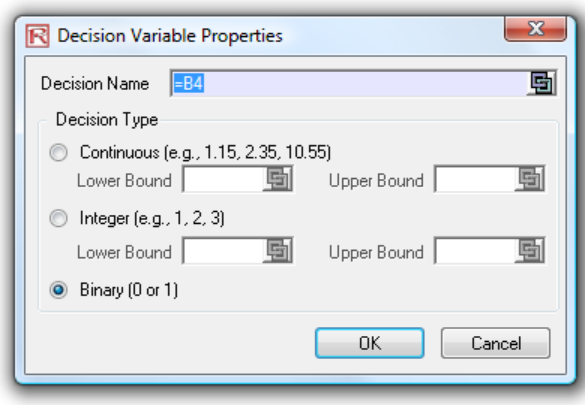
- The final step in optimization is to set the objective function and start the optimization by selecting cell *C19* and *Risk Simulator | Optimization | Set Objective*. Then run the optimization using *Risk Simulator | Optimization | Run Optimization* and selecting the optimization of choice (*Static Optimization, Dynamic Optimization*, or *Stochastic Optimization*). To get started, select *Static Optimization*. Check to make sure that the *objective* cell is either *the Sharpe ratio* or *portfolio returns* to risk ratio and select Maximize. You can now review the decision variables and constraints if required, or click *OK* to run the static optimization.

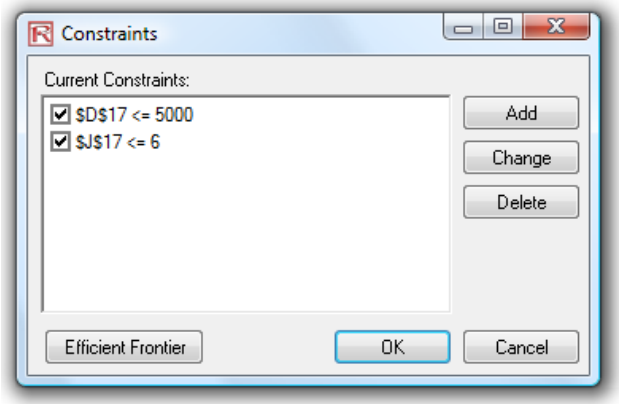
Figure 4.5 shows the screen shots of these procedural steps. You can add simulation assumptions on the model's ENPV and risk (columns C and E), and apply the dynamic optimization and stochastic optimization for additional practice.

	ENPV	Cost	Risk $	Risk %	Return to Risk Ratio	Profitability Index	Selection
Project 1	$458.00	$1,732.44	$54.96	12.00%	8.33	1.26	1.0000
Project 2	$1,954.00	$859.00	$1,914.92	98.00%	1.02	3.27	1.0000
Project 3	$1,599.00	$1,845.00	$1,551.03	97.00%	1.03	1.87	1.0000
Project 4	$2,251.00	$1,645.00	$1,012.95	45.00%	2.22	2.37	1.0000
Project 5	$849.00	$458.00	$925.41	109.00%	0.92	2.85	1.0000
Project 6	$758.00	$52.00	$560.92	74.00%	1.35	15.58	1.0000
Project 7	$2,845.00	$758.00	$5,633.10	198.00%	0.51	4.75	1.0000
Project 8	$1,235.00	$115.00	$926.25	75.00%	1.33	11.74	1.0000
Project 9	$1,945.00	$125.00	$2,100.60	108.00%	0.93	16.56	1.0000
Project 10	$2,250.00	$458.00	$1,912.50	85.00%	1.18	5.91	1.0000
Project 11	$549.00	$45.00	$263.52	48.00%	2.08	13.20	1.0000
Project 12	$525.00	$105.00	$309.75	59.00%	1.69	6.00	1.0000
Total	$17,218.00	$8,197.44	$7,007	40.70%			12.00
Goal:	MAX	< =$5000					<=6
Sharpe Ratio	2.4573						

ENPV is the expected NPV of each credit line or project, while Cost can be the total cost of administration as well as required capital holdings to cover the credit line, and Risk is the Coefficient of Variation of the credit line's ENPV.

Figure 4.4 – Discrete Integer Optimization Model





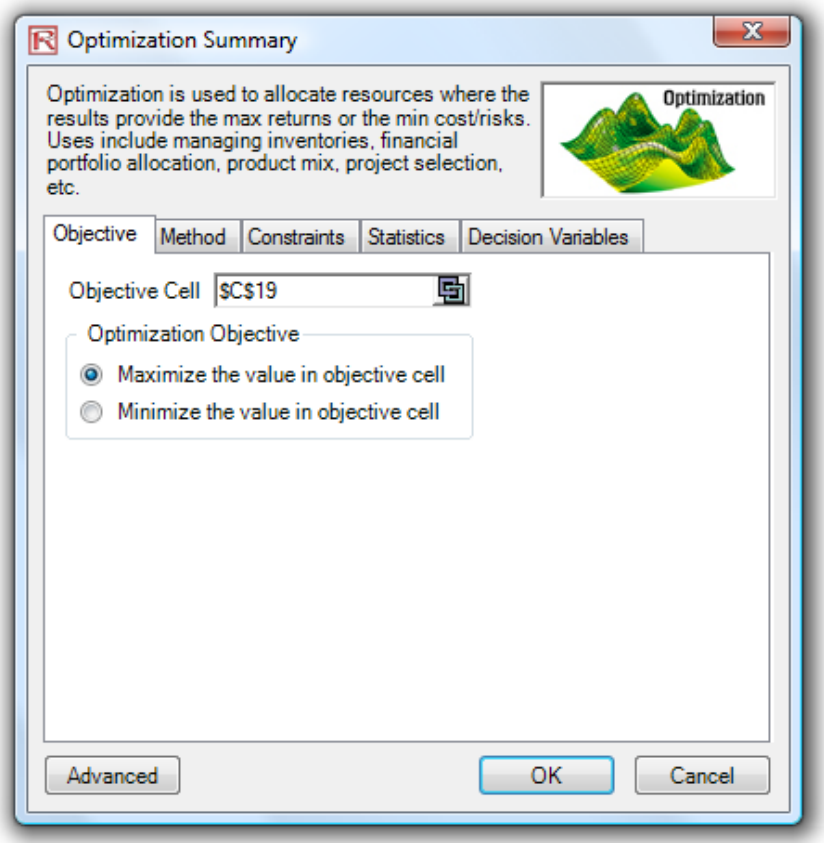


Figure 4.5 – Running Discrete Integer Optimization in Risk Simulator

Results Interpretation

Figure 4.6 shows a sample optimal selection of projects that maximizes the Sharpe ratio. In contrast, one can always maximize total revenues, but, as before, this is a trivial process and simply involves choosing the highest returning project and going down the list until you run out of money or exceed the budget constraint. Doing so will yield theoretically undesirable projects as the highest yielding projects typically hold higher risks. Now, if desired, you can replicate the optimization using a stochastic or dynamic optimization by adding assumptions in the ENPV and/or cost, and/or risk values.

For additional hands-on examples of optimization in action, see the case study in Chapter 11 on Integrated Risk Management in the book, Real Options Analysis: Tools and Techniques, *Second Edition* (Wiley Finance, 2010), by Dr. Johnathan Mun. That case study illustrates how an efficient frontier can be generated and how forecasting, simulation, optimization, and real options can be combined into a seamless analytical process.

	ENPV	Cost	Risk $	Risk %	Return to Risk Ratio	Profitability Index	Selection
Project 1	$458.00	$1,732.44	$54.96	12.00%	8.33	1.26	1.0000
Project 2	$1,954.00	$859.00	$1,914.92	98.00%	1.02	3.27	0.0000
Project 3	$1,599.00	$1,845.00	$1,551.03	97.00%	1.03	1.87	0.0000
Project 4	$2,251.00	$1,645.00	$1,012.95	45.00%	2.22	2.37	1.0000
Project 5	$849.00	$458.00	$925.41	109.00%	0.92	2.85	0.0000
Project 6	$758.00	$52.00	$560.92	74.00%	1.35	15.58	1.0000
Project 7	$2,845.00	$758.00	$5,633.10	198.00%	0.51	4.75	0.0000
Project 8	$1,235.00	$115.00	$926.25	75.00%	1.33	11.74	1.0000
Project 9	$1,945.00	$125.00	$2,100.60	108.00%	0.93	16.56	0.0000
Project 10	$2,250.00	$458.00	$1,912.50	85.00%	1.18	5.91	0.0000
Project 11	$549.00	$45.00	$263.52	48.00%	2.08	13.20	1.0000
Project 12	$525.00	$105.00	$309.75	59.00%	1.69	6.00	1.0000
Total	$5,776.00	$3,694.44	$1,539	26.64%			6.00
Goal:	MAX	<=$5000					<=6
Sharpe Ratio	3.7543						

ENPV is the expected NPV of each credit line or project, while Cost can be the total cost of administration as well as required capital holdings to cover the credit line, and Risk is the Coefficient of Variation of the credit line's ENPV.

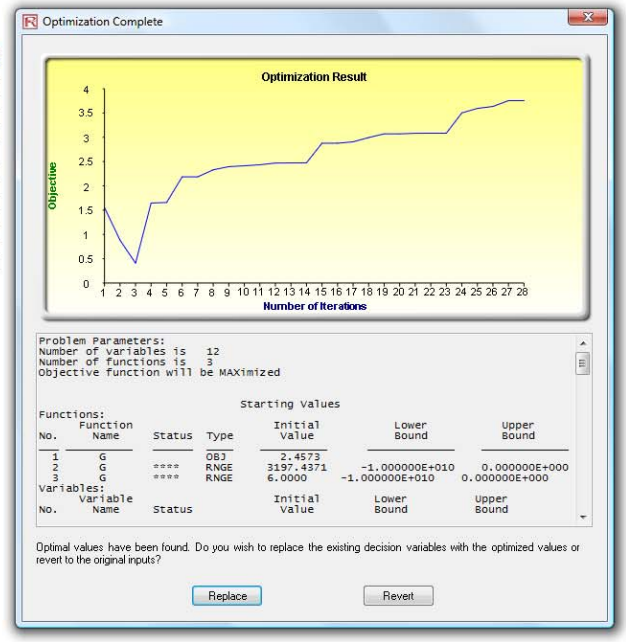


Figure 4.6 – Optimal Selection of Projects That Maximizes the Sharpe Ratio

4.4 Efficient Frontier and Advanced Optimization Settings

The middle graphic in Figure 4.5 shows the constraints set for the example optimization. Within this function, if you click on the Efficient Frontier button after you have set some constraints, you can make the constraints changing. That is, each of the constraints can be created to step through between some maximum and minimum value. As an example, the constraint in cell *J17 <= 6* can be set to *run between 4 and 8* (Figure 4.7). Thus, five optimizations will be run, each with the following constraints: J17 <= 4, J17 <= 5, J17 <= 6, J17 <= 7, and J17 <= 8. The optimal results will then be plotted as an efficient frontier and the report will be generated (Figure 4.8). Specifically, here are the steps required to create a changing constraint:

- In an optimization model (i.e., a model with Objective, Decision Variables, and Constraints already set up), click on *Risk Simulator | Optimization | Constraints* and click on *Efficient Frontier*.
- Select the constraint you want to change or step (e.g., J17), enter in the parameters for *Min, Max,* and *Step Size* (Figure 4.7), click *ADD*, and then click *OK* and *OK* again. You should *deselect D17 <= 5000* constraint before running.
- Run Optimization as usual (*Risk Simulator | Optimization | Run Optimization*). You can choose static, dynamic, or stochastic.
- The results will be shown as a user interface (Figure 4.8). Click on Create Report to generate a report worksheet with all the details of the optimization runs.

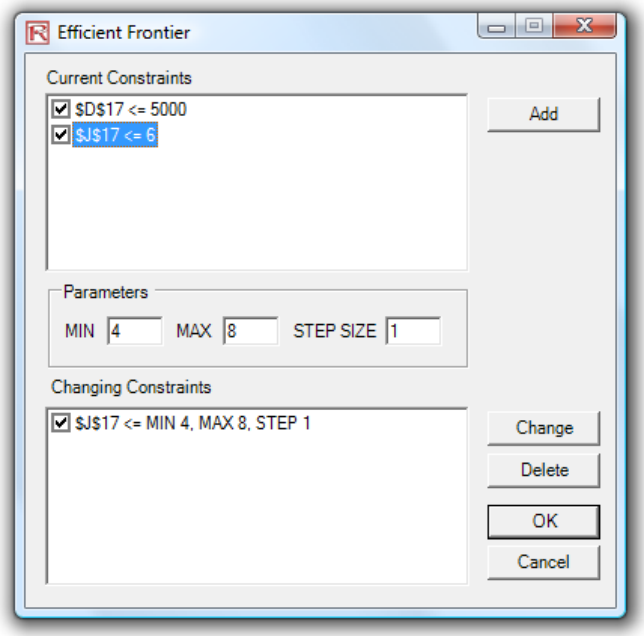


Figure 4.7 – Generating Changing Constraints in an Efficient Frontier

Efficient Frontier

Problem Parameters:

Number of variables	12
Number of functions	3
Objective function will be	Maximized

STEP1, D17 <= 5000, J17 <= 4

Functions

Starting Values

No.	Function Name	Status	Type	Initial Value	Lower Bound	Upper Bound
1	G		OBJ	2.45726		
2	G	****	RNGE	3197.43710	-1E+10	0
3	G	****	RNGE	8.00000	-1E+10	0

Final Results

No.	Function Name	Initial Value	Final Value
1	G	2.45726	3.46137
2	G	3197.43710	-1472.56292
3	G	8.00000	0.00000

Variables

Starting Values

No.	Variable Name	Status	Initial Value	Lower Bound	Upper Bound
1	X	UL	1.00000	0	1
2	X	UL	1.00000	0	1
3	X	UL	1.00000	0	1
4	X	UL	1.00000	0	1
5	X	UL	1.00000	0	1
6	X	UL	1.00000	0	1
7	X	UL	1.00000	0	1
8	X	UL	1.00000	0	1
9	X	UL	1.00000	0	1
10	X	UL	1.00000	0	1
11	X	UL	1.00000	0	1
12	X	UL	1.00000	0	1

Final Results

No.	Variable Name	Initial Value	Final Value
1	X	1.00000	1.00000
2	X	1.00000	0.00000
3	X	1.00000	0.00000
4	X	1.00000	1.00000
5	X	1.00000	0.00000
6	X	1.00000	0.00000
7	X	1.00000	0.00000
8	X	1.00000	0.00000
9	X	1.00000	0.00000
10	X	1.00000	0.00000
11	X	1.00000	1.00000
12	X	1.00000	1.00000

No.	Objective Function	Binding Constrs	Super Basics	Infeas Constr	Norm of Red. Grad	Hessian Cond.No.	Step Size	Degen Step
1	3205.43710	0	12	2	0.57590	1	0	
2	3.55285	0	11	1	0.28146	1	1	
3	2.88211	0	10	1	0.34697	1	0.061	

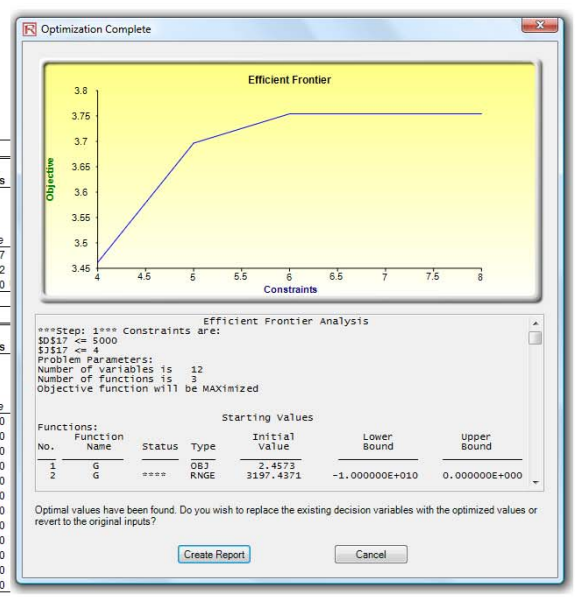


Figure 4.8 – Efficient Frontier Results

4.5 Stochastic Optimization

This example illustrates the application of stochastic optimization using a sample model with four asset classes each with different risk and return characteristics. The idea here is to find the best portfolio allocation such that the portfolio's bang for the buck, or returns to risk ratio, is maximized. That is, the goal is to allocate 100% of an individual's investment among several different asset classes (e.g., different types of mutual funds or investment styles: growth, value, aggressive growth, income, global, index, contrarian, momentum, etc.). This model is different from others in that there exists several simulation assumptions (risk and return values for each asset in columns C and D), as seen in Figure 4.9.

A simulation is run, then optimization is executed, and the entire process is repeated multiple times to obtain distributions of each decision variable. The entire analysis can be automated using Stochastic Optimization. To run an optimization, several key specifications on the model have to be identified first:

Objective:	*Maximize Return to Risk Ratio (C12)*
Decision Variables:	*Allocation Weights (E6:E9)*
Restrictions on Decision Variables:	*Minimum and Maximum Required (F6:G9)*
Constraints:	*Portfolio Total Allocation Weights 100%*
	(E11 is set to 100%)
Simulation Assumptions:	*Return and Risk Values (C6:D9)*

The model shows the various asset classes. Each asset class has its own set of annualized returns and annualized volatilities. These return and risk measures are annualized values such that they can be consistently compared across different asset classes. Returns are computed using the geometric average of the relative returns, while the risks are computed using the logarithmic relative stock returns approach.

In Figure 4.9, column E (Allocation Weights) holds the decision variables, which are the variables that need to be tweaked and tested such that the total weight is constrained at 100% (cell E11). Typically, to start the optimization, we set these cells to a uniform value. In this case, cells E6 to E9 are set at 25% each. In addition, each decision variable may have specific restrictions in its allowed range. In this example, the lower and upper allocations allowed are 10% and 40%, as seen in columns F and G. This setting means that each asset class may have its own allocation boundaries.

Next, column H shows the return to risk ratio, which is simply the return percentage divided by the risk percentage for each asset, where the higher this value, the higher the bang for the buck. The remaining parts of the model show the individual asset class rankings by returns, risk, return to risk ratio, and allocation. In other words, these rankings show at a glance which asset class has the lowest risk, or the highest return, and so forth.

	A	B	C	D	E	F	G	H
1								
2								
3			ASSET ALLOCATION OPTIMIZATION MODEL					
4								
5		Asset Class Description	Annualized Returns	Volatility Risk	Allocation Weights	Required Minimum Allocation	Required Maximum Allocation	Return to Risk Ratio
6		Asset 1	10.60%	12.41%	25.00%	10.00%	40.00%	0.8544
7		Asset 2	11.21%	16.16%	25.00%	10.00%	40.00%	0.6937
8		Asset 3	10.61%	15.93%	25.00%	10.00%	40.00%	0.6660
9		Asset 4	10.52%	12.40%	25.00%	10.00%	40.00%	0.8480
10								
11		***Portfolio Total***	***10.7356%***	***7.17%***	***100.00%***			
12		***Return to Risk Ratio***	***1.4970***					

Figure 4.9 – Asset Allocation Model Ready for Stochastic Optimization

Procedure

To run this model, simply click on *Risk Simulator | Optimization | Run Optimization*. Alternatively, and for practice, you can set up the model using the following steps illustrated in Figure 4.10:

1. Start a new profile (*Risk Simulator | New Profile*). For stochastic optimization, set distributional assumptions on the risk and returns for each asset class. That is, select cell C6, set an assumption (*Risk Simulator | Set Input Assumption*), and designate your own assumption as required. Repeat for cells C7 to D9.

2. Select cell E6, and define the decision variable (*Risk Simulator | Optimization | Set Decision* or click on the *Set Decision D* icon) and make it a Continuous Variable. Then link the decision variable's name and minimum/maximum required to the relevant cells (*B6*, *F6*, *G6*).

3. Then use Risk Simulator's copy on cell *E6*, select cells *E7* to *E9*, and use Risk Simulator's paste (*Risk Simulator | Copy Parameter* and *Risk Simulator | Paste Parameter* or use the copy and paste icons). Remember not to use Excel's regular copy and paste functions.

4. Next, set up the optimization's constraints by selecting *Risk Simulator | Optimization | Constraints*, selecting *ADD*, and selecting the cell *E11* and making it equal *100%* (total allocation, and do not forget the % sign).

5. Select cell *C12*, the objective to be maximized, and make it the objective: *Risk Simulator | Optimization | Set Objective* or click on the *O* icon. Run the optimization by going to *Risk Simulator | Optimization | Run Optimization*. Review the different tabs to make sure that all the required inputs in steps 2 and 3 are correct. Select *Stochastic Optimization* and let it run for *500* trials repeated *20* times. Click *OK* when the simulation completes and a detailed stochastic optimization report will be generated along with forecast charts of the decision variables.

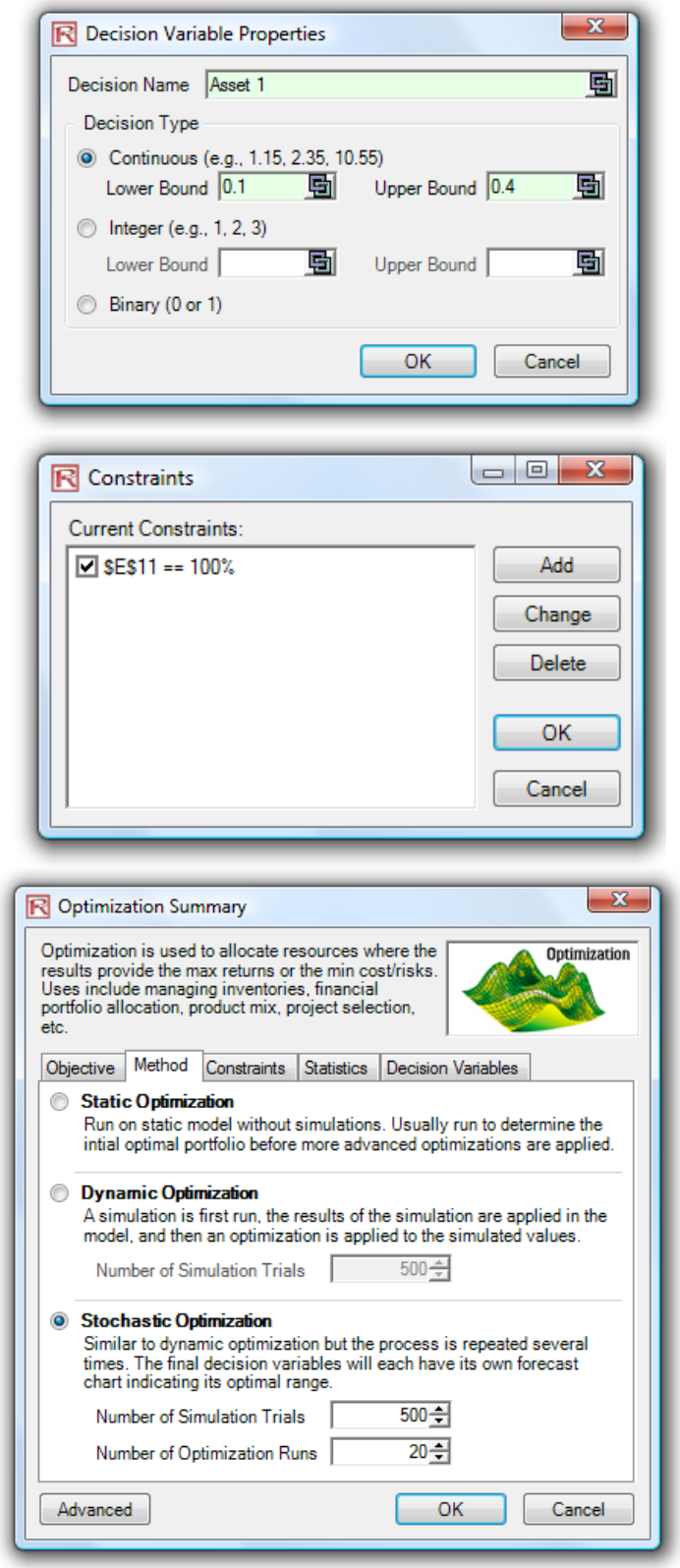


Figure 4.10 – Setting Up the Stochastic Optimization Problem

Results Interpretation

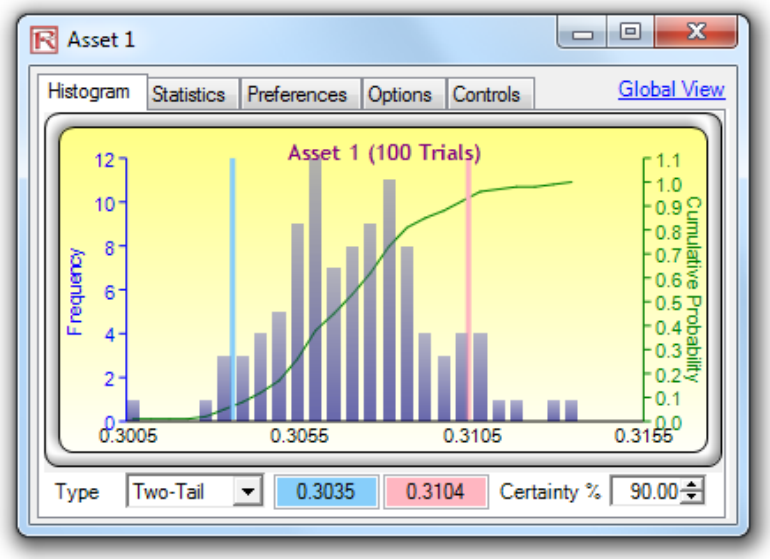
Stochastic optimization is performed when a simulation is run first and then the optimization is run. Then the whole analysis is repeated multiple times. As shown in Figure 4.11 for the example optimization, the result is a distribution of each decision variable rather than a single-point estimate. This means that instead of saying you should invest 30.69% in Asset 1, the results show that the optimal decision is to invest between 30.35% and 31.04% as long as the total portfolio sums to 100%. This way, the results provide management or decision makers a range of flexibility in the optimal decisions while accounting for the risks and uncertainties in the inputs.

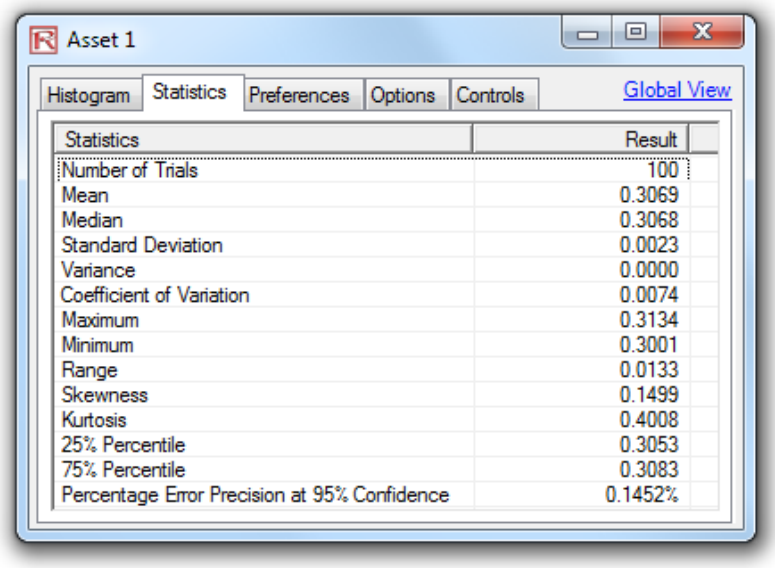
Super Speed Simulation with Optimization

You can also run stochastic optimization with super speed simulation. To do this, first reset the optimization by resetting all four decision variables back to 25%. Next, Run Optimization, click on the Advanced button (Figure 4.10), and select the checkbox for Run Super Speed Simulation. Then, in the run optimization user interface, select Stochastic Optimization on the Method tab and set it to run 500 trials and 20 optimization runs, and click OK. This approach will integrate the super speed simulation with optimization. Notice how much faster the stochastic optimization runs. You can now quickly rerun the optimization with a higher number of simulation trials.

Simulation Statistics for Stochastic and Dynamic Optimization

Notice that if there are input simulation assumptions in the optimization model (i.e., these input assumptions are required in order to run the dynamic or stochastic optimization routines), the Statistics tab is now populated in the Run Optimization user interface. You can select from the drop-down list the statistics you want, such as average, standard deviation, coefficient of variation, conditional mean, conditional variance, a specific percentile, and so forth. This means that if you run a stochastic optimization, a simulation of thousands of trials will first run, then the selected statistic will be computed and this value will be temporarily placed in the simulation assumption cell, then an optimization will be run based on this statistic, and then the entire process is repeated multiple times. This method is important and useful for banking applications in computing conditional Value at Risk, or conditional VaR.





Statistics	Result
Number of Trials	100
Mean	0.3069
Median	0.3068
Standard Deviation	0.0023
Variance	0.0000
Coefficient of Variation	0.0074
Maximum	0.3134
Minimum	0.3001
Range	0.0133
Skewness	0.1499
Kurtosis	0.4008
25% Percentile	0.3053
75% Percentile	0.3083
Percentage Error Precision at 95% Confidence	0.1452%

Figure 4.11 – Simulated Results from the Stochastic Optimization Approach

5

5. RISK SIMULATOR ANALYTICAL TOOLS

This chapter covers Risk Simulator's analytical tools, providing detailed discussions of the applicability of each tool and through example applications, complete with step-by-step illustrations. These tools are very valuable to analysts working in the realm of risk analysis.

5.1 Tornado and Sensitivity Tools in Simulation

Theory

Tornado analysis is a powerful simulation tool that captures the static impacts of each variable on the outcome of the model. That is, the tool automatically perturbs each variable in the model a preset amount, captures the fluctuation on the model's forecast or final result, and lists the resulting perturbations ranked from the most significant to the least. Figures 5.1 through 5.6 illustrate the application of a tornado analysis. For instance, Figure 5.1 is a sample discounted cash flow model where the input assumptions in the model are shown. The question is what are the critical success drivers that affect the model's output the most? That is, what really drives the net present value of $96.63 or which input variable impacts this value the most?

The tornado chart tool can be accessed through *Risk Simulator | Tools | Tornado Analysis*. To follow along the first example, open the Tornado and Sensitivity Charts (Linear) file in the examples folder. Figure 5.2 shows this sample model where cell G6 containing the net present value is chosen as the target result to be analyzed. The target cell's precedents in the model are used in creating the tornado chart. Precedents are all the input and intermediate variables that affect the outcome of the model. For instance, if the model consists of A = B + C, and where C = D + E, then B, D, and E are the precedents for A (C is not a precedent as it is only an intermediate calculated value). Figure 5.2 also shows the testing range of each precedent variable used to estimate the target result. If the precedent variables are simple inputs, then the testing range will be a simple perturbation based on the range chosen (e.g., the default is ±10%). Each precedent variable can be perturbed at different percentages if required. A wider range is important as it is better able to test extreme values rather than smaller perturbations around the expected values. In certain circumstances, extreme values may have a larger, smaller, or unbalanced impact (e.g., nonlinearities may occur where increasing or decreasing economies of scale and scope creep in for larger or smaller values of a variable) and only a wider range will capture this nonlinear impact.

Procedure

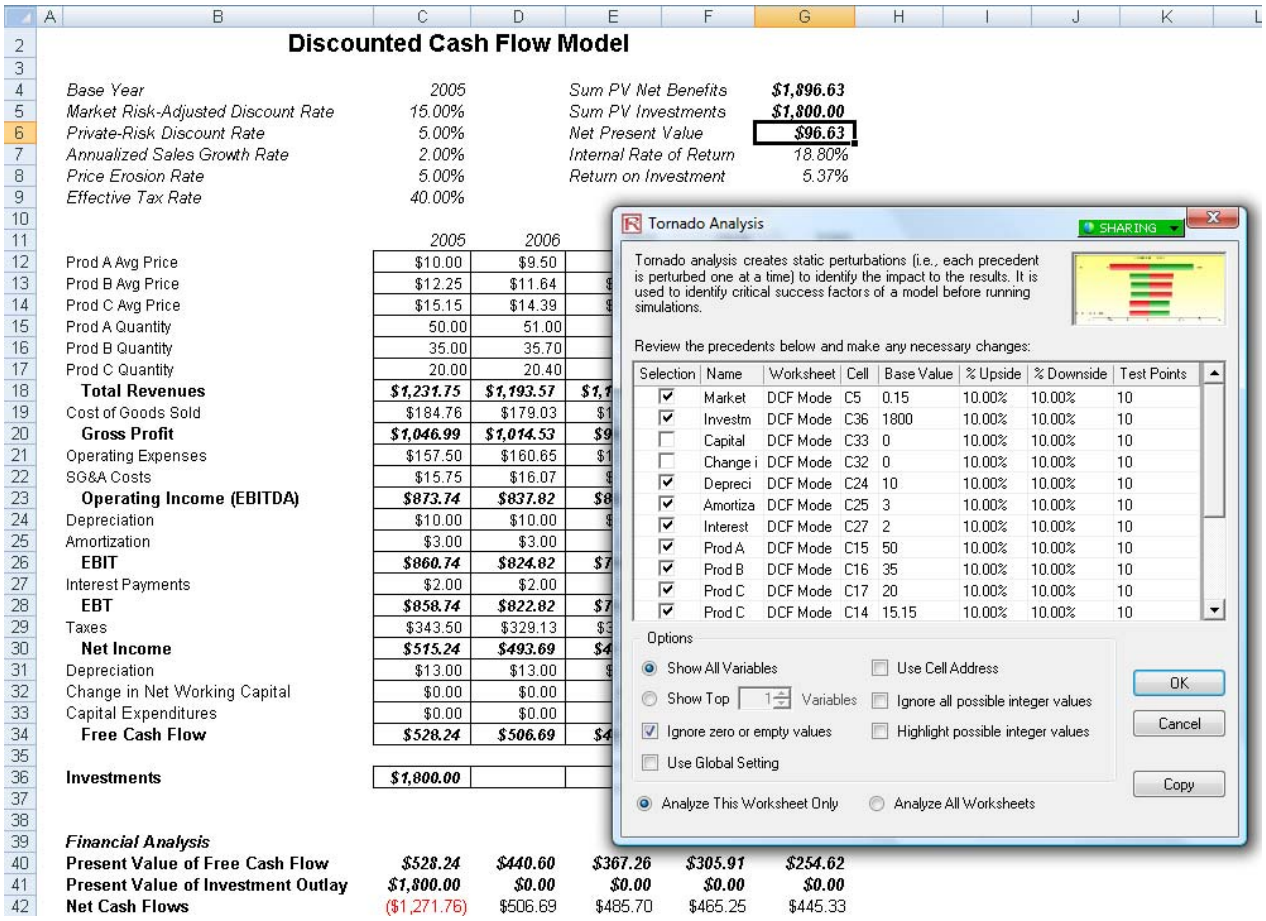
- Select the single output cell (i.e., a cell with a function or equation) in an Excel model (e.g., cell *G6* is selected in our example).
- Select *Risk Simulator | Tools | Tornado Analysis*.
- Review the precedents and rename them as needed (renaming the precedents to shorter names allows a more visually pleasing tornado and spider chart), and click *OK*.

Discounted Cash Flow Model

Base Year	*2005*	*Sum PV Net Benefits*	***$1,896.63***
Market Risk-Adjusted Discount Rate	*15.00%*	*Sum PV Investments*	***$1,800.00***
Private-Risk Discount Rate	*5.00%*	*Net Present Value*	***$96.63***
Annualized Sales Growth Rate	*2.00%*	*Internal Rate of Return*	*18.80%*
Price Erosion Rate	*5.00%*	*Return on Investment*	*5.37%*
Effective Tax Rate	*40.00%*		

	2005	*2006*	*2007*	*2008*	*2009*
Product A Avg Price/Unit	$10.00	$9.50	$9.03	$8.57	$8.15
Product B Avg Price/Unit	$12.25	$11.64	$11.06	$10.50	$9.98
Product C Avg Price/Unit	$15.15	$14.39	$13.67	$12.99	$12.34
Product A Sale Quantity ('000s)	50.00	51.00	52.02	53.06	54.12
Product B Sale Quantity ('000s)	35.00	35.70	36.41	37.14	37.89
Product C Sale Quantity ('000s)	20.00	20.40	20.81	21.22	21.65
Total Revenues	***$1,231.75***	***$1,193.57***	***$1,156.57***	***$1,120.71***	***$1,085.97***
Direct Cost of Goods Sold	$184.76	$179.03	$173.48	$168.11	$162.90
Gross Profit	***$1,046.99***	***$1,014.53***	***$983.08***	***$952.60***	***$923.07***
Operating Expenses	$157.50	$160.65	$163.86	$167.14	$170.48
Sales, General and Admin. Costs	$15.75	$16.07	$16.39	$16.71	$17.05
Operating Income (EBITDA)	***$873.74***	***$837.82***	***$802.83***	***$768.75***	***$735.54***
Depreciation	$10.00	$10.00	$10.00	$10.00	$10.00
Amortization	$3.00	$3.00	$3.00	$3.00	$3.00
EBIT	***$860.74***	***$824.82***	***$789.83***	***$755.75***	***$722.54***
Interest Payments	$2.00	$2.00	$2.00	$2.00	$2.00
EBT	***$858.74***	***$822.82***	***$787.83***	***$753.75***	***$720.54***
Taxes	$343.50	$329.13	$315.13	$301.50	$288.22
Net Income	***$515.24***	***$493.69***	***$472.70***	***$452.25***	***$432.33***
Noncash: Depreciation Amortization	$13.00	$13.00	$13.00	$13.00	$13.00
Noncash: Change in Net Working Capital	$0.00	$0.00	$0.00	$0.00	$0.00
Noncash: Capital Expenditures	$0.00	$0.00	$0.00	$0.00	$0.00
Free Cash Flow	***$528.24***	***$506.69***	***$485.70***	***$465.25***	***$445.33***
Investment Outlay	***$1,800.00***				
Financial Analysis					
Present Value of Free Cash Flow	***$528.24***	***$440.60***	***$367.26***	***$305.91***	***$254.62***
Present Value of Investment Outlay	***$1,800.00***	***$0.00***	***$0.00***	***$0.00***	***$0.00***
Net Cash Flows	($1,271.76)	$506.69	$485.70	$465.25	$445.33

Figure 5.1 – Sample Model



Discounted Cash Flow Model

Base Year	*2005*	*Sum PV Net Benefits*	***$1,896.63***
Market Risk-Adjusted Discount Rate	*15.00%*	*Sum PV Investments*	***$1,800.00***
Private-Risk Discount Rate	*5.00%*	*Net Present Value*	***$96.63***
Annualized Sales Growth Rate	*2.00%*	*Internal Rate of Return*	*18.80%*
Price Erosion Rate	*5.00%*	*Return on Investment*	*5.37%*
Effective Tax Rate	*40.00%*		

	2005	*2006*			
Prod A Avg Price	$10.00	$9.50			
Prod B Avg Price	$12.25	$11.64			
Prod C Avg Price	$15.15	$14.39			
Prod A Quantity	50.00	51.00			
Prod B Quantity	35.00	35.70			
Prod C Quantity	20.00	20.40			
Total Revenues	***$1,231.75***	***$1,193.57***			
Cost of Goods Sold	$184.76	$179.03			
Gross Profit	***$1,046.99***	***$1,014.53***			
Operating Expenses	$157.50	$160.65			
SG&A Costs	$15.75	$16.07			
Operating Income (EBITDA)	***$873.74***	***$837.82***			
Depreciation	$10.00	$10.00			
Amortization	$3.00	$3.00			
EBIT	***$860.74***	***$824.82***			
Interest Payments	$2.00	$2.00			
EBT	***$858.74***	***$822.82***			
Taxes	$343.50	$329.13			
Net Income	***$515.24***	***$493.69***			
Depreciation	$13.00	$13.00			
Change in Net Working Capital	$0.00	$0.00			
Capital Expenditures	$0.00	$0.00			
Free Cash Flow	***$528.24***	***$506.69***			
Investments	***$1,800.00***				
Financial Analysis					
Present Value of Free Cash Flow	***$528.24***	***$440.60***	***$367.26***	***$305.91***	***$254.62***
Present Value of Investment Outlay	***$1,800.00***	***$0.00***	***$0.00***	***$0.00***	***$0.00***
Net Cash Flows	($1,271.76)	$506.69	$485.70	$465.25	$445.33

Tornado Analysis

SHARING

Tornado analysis creates static perturbations (i.e., each precedent is perturbed one at a time) to identify the impact to the results. It is used to identify critical success factors of a model before running simulations.

Review the precedents below and make any necessary changes:

Selection	Name	Worksheet	Cell	Base Value	% Upside	% Downside	Test Points
☑	Market	DCF Mode	C5	0.15	10.00%	10.00%	10
☑	Investm	DCF Mode	C36	1800	10.00%	10.00%	10
☐	Capital	DCF Mode	C33	0	10.00%	10.00%	10
☐	Change i	DCF Mode	C32	0	10.00%	10.00%	10
☑	Depreci	DCF Mode	C24	10	10.00%	10.00%	10
☑	Amortiza	DCF Mode	C25	3	10.00%	10.00%	10
☑	Interest	DCF Mode	C27	2	10.00%	10.00%	10
☑	Prod A	DCF Mode	C15	50	10.00%	10.00%	10
☑	Prod B	DCF Mode	C16	35	10.00%	10.00%	10
☑	Prod C	DCF Mode	C17	20	10.00%	10.00%	10
☑	Prod C	DCF Mode	C14	15.15	10.00%	10.00%	10

Options

Show All Variables — Use Cell Address

Show Top 1 Variables — Ignore all possible integer values

Ignore zero or empty values — Highlight possible integer values

Use Global Setting

Analyze This Worksheet Only — Analyze All Worksheets

OK

Cancel

Copy

Figure 5.2 – Running Tornado Analysis

Results Interpretation

Figure 5.3 shows the resulting tornado analysis report, which indicates that capital investment has the largest impact on net present value, followed by tax rate, average sale price, quantity demanded of the product lines, and so forth. The report contains four distinct elements:

A statistical summary listing the procedure performed.

A sensitivity table (Figure 5.4) showing the starting NPV base value of 96.63 and how each input is changed (e.g., Investment is changed from $1,800 to $1,980 on the upside with a +10% swing, and from $1,800 to $1,620 on the downside with a –10% swing. The resulting upside and downside values on NPV is –$83.37 and $276.63, with a total change of $360, making investment the variable with the highest impact on NPV.) The precedent variables are ranked from the highest impact to the lowest impact.

A spider chart (Figure 5.5) illustrating the effects graphically. The y-axis is the NPV target value while the x-axis depicts the percentage change on each of the precedent values (the central point is the base case value at 96.63 at 0% change from the base value of each precedent). A positively sloped line indicates a positive relationship or effect, while negatively sloped lines indicate a negative relationship (e.g., Investment is negatively sloped, which means that the higher the investment level, the lower the NPV). The absolute value of the slope indicates the

magnitude of the effect (a steep line indicates a higher impact on the NPV y-axis given a change in the precedent x-axis).

A tornado chart illustrating the effects in another graphical manner, where the highest impacting precedent is listed first. The x-axis is the NPV value, with the center of the chart being the base case condition. Green bars in the chart indicate a positive effect, while red bars indicate a negative effect. Therefore, for investments, the red bar on the right side indicates a negative effect of investment on higher NPV—in other words, capital investment and NPV are negatively correlated. The opposite is true for price and quantity of products A to C (their green bars are on the right side of the chart).

Tornado and Spider Charts

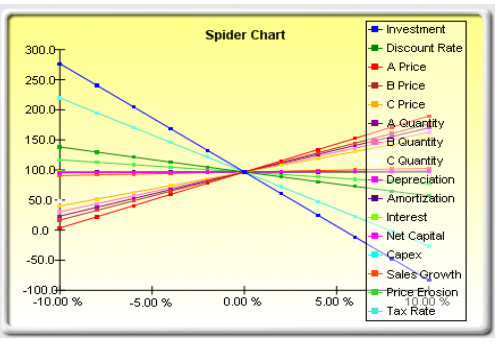
Statistical Summary

One of the powerful simulation tools is the tornado chart—it captures the static impacts of each variable on the outcome of the model. That is, the tool automatically perturbs each precedent variable in the model a user-specified preset amount, captures the fluctuation on the model's forecast or final result, and lists the resulting perturbations ranked from the most significant to the least. Precedents are all the input and intermediate variables that affect the outcome of the model. For instance, if the model consists of A = B + C, where C = D + E, then B, D, and E are the precedents for A (C is not a precedent as it is only an intermediate calculated value). The range and number of values perturbed is user-specified and can be set to test extreme values rather than smaller perturbations around the expected values. In certain circumstances, extreme values may have a larger, smaller, or unbalanced impact (e.g., nonlinearities may occur where increasing or decreasing economies of scale and scope creep occurs for larger or smaller values of a variable) and only a wider range will capture this nonlinear impact.

A tornado chart lists all the inputs that drive the model, starting from the input variable that has the most effect on the results. The chart is obtained by perturbing each precedent input at some consistent range (e.g., ±10% from the base case) one at a time, and comparing their results to the base case. A spider chart looks like a spider with a central body and its many legs protruding. The positively sloped lines indicate a positive relationship, while a negatively sloped line indicates a negative relationship. Further, spider charts can be used to visualize linear and nonlinear relationships. The tornado and spider charts help identify the critical success factors of an output cell in order to identify the inputs to simulate. The identified critical variables that are uncertain are the ones that should be simulated. Do not waste time simulating variables that are neither uncertain nor have little impact on the results.

Result

	Base Value: 96.6261638553219			Input Changes		
Precedent Cell	Output Downside	Output Upside	Effective Range	Input Downside	Input Upside	Base Case Value
Investment	$276.63	($83.37)	360.00	$1,620.00	$1,980.00	$1,800.00
Tax Rate	$219.73	($26.47)	246.20	36.00%	44.00%	40.00%
A Price	$3.43	$189.83	186.40	$9.00	$11.00	$10.00
B Price	$16.71	$176.55	159.84	$11.03	$13.48	$12.25
A Quantity	$23.18	$170.07	146.90	45.00	55.00	50.00
B Quantity	$30.53	$162.72	132.19	31.50	38.50	35.00
C Price	$40.15	$153.11	112.96	$13.64	$16.67	$15.15
C Quantity	$48.05	$145.20	97.16	18.00	22.00	20.00
Discount Rate	$138.24	$57.03	81.21	13.50%	16.50%	15.00%
Price Erosion	$116.80	$76.64	40.16	4.50%	5.50%	5.00%
Sales Growth	$90.59	$102.69	12.10	1.80%	2.20%	2.00%
Depreciation	$95.08	$98.17	3.08	$9.00	$11.00	$10.00
Interest	$97.09	$96.16	0.93	$1.80	$2.20	$2.00
Amortization	$96.16	$97.09	0.93	$2.70	$3.30	$3.00
Capex	$96.63	$96.63	0.00	$0.00	$0.00	$0.00
Net Capital	$96.63	$96.63	0.00	$0.00	$0.00	$0.00



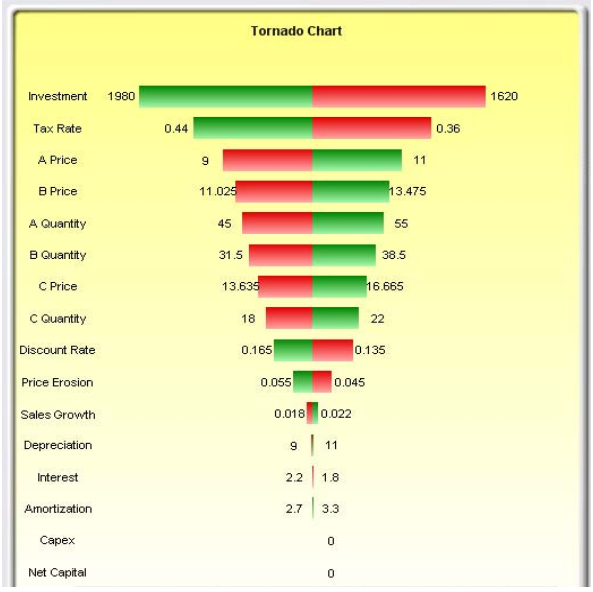


Figure 5.3 – Tornado Analysis Report

Notes

Remember that tornado analysis is a static sensitivity analysis applied on each input variable in the model—that is, each variable is perturbed individually and the resulting effects are tabulated. This approach makes tornado analysis a key component to execute before running a simulation. One of the very first steps in risk analysis is capturing and identifying the most important impact drivers in the model. The next step is to identify which of these important impact drivers are uncertain. These uncertain impact drivers are the critical success drivers of a project, where the results of the model depend on these critical success drivers. These variables are the ones that should be simulated. Do not waste time simulating variables that are neither uncertain nor have little impact on the results. Tornado charts assist in identifying these critical success drivers quickly and easily. Following this example, it might be that price and quantity should be simulated, assuming that the required investment and effective tax rate are both known in advance and unchanging.

	Base Value: 96.6261638553219			Input Changes		
Precedent Cell	Output Downside	Output Upside	Effective Range	Input Downside	Input Upside	Base Case Value
Investment	$276.63	($83.37)	360.00	$1,620.00	$1,980.00	$1,800.00
Tax Rate	$219.73	($26.47)	246.20	36.00%	44.00%	40.00%
A Price	$3.43	$189.83	186.40	$9.00	$11.00	$10.00
B Price	$16.71	$176.55	159.84	$11.03	$13.48	$12.25
A Quantity	$23.18	$170.07	146.90	45.00	55.00	50.00
B Quantity	$30.53	$162.72	132.19	31.50	38.50	35.00
C Price	$40.15	$153.11	112.96	$13.64	$16.67	$15.15
C Quantity	$48.05	$145.20	97.16	18.00	22.00	20.00
Discount Rate	$138.24	$57.03	81.21	13.50%	16.50%	15.00%
Price Erosion	$116.80	$76.64	40.16	4.50%	5.50%	5.00%
Sales Growth	$90.59	$102.69	12.10	1.80%	2.20%	2.00%
Depreciation	$95.08	$98.17	3.08	$9.00	$11.00	$10.00
Interest	$97.09	$96.16	0.93	$1.80	$2.20	$2.00
Amortization	$96.16	$97.09	0.93	$2.70	$3.30	$3.00
Capex	$96.63	$96.63	0.00	$0.00	$0.00	$0.00
Net Capital	$96.63	$96.63	0.00	$0.00	$0.00	$0.00

Figure 5.4 – Sensitivity Table

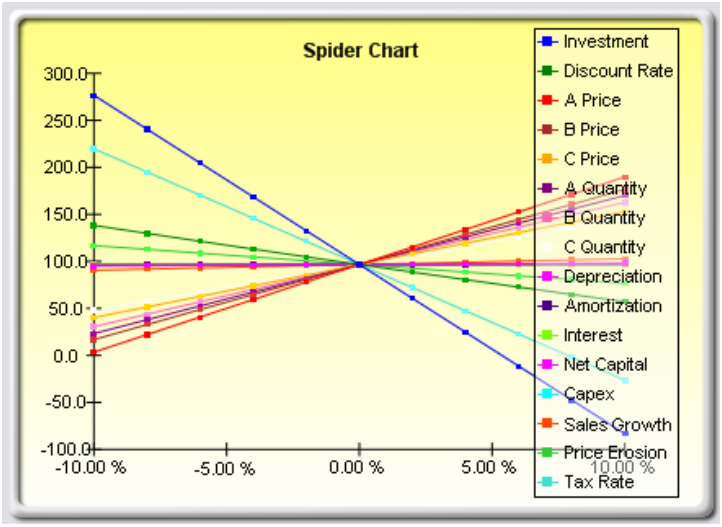


Figure 5.5 – Spider Chart

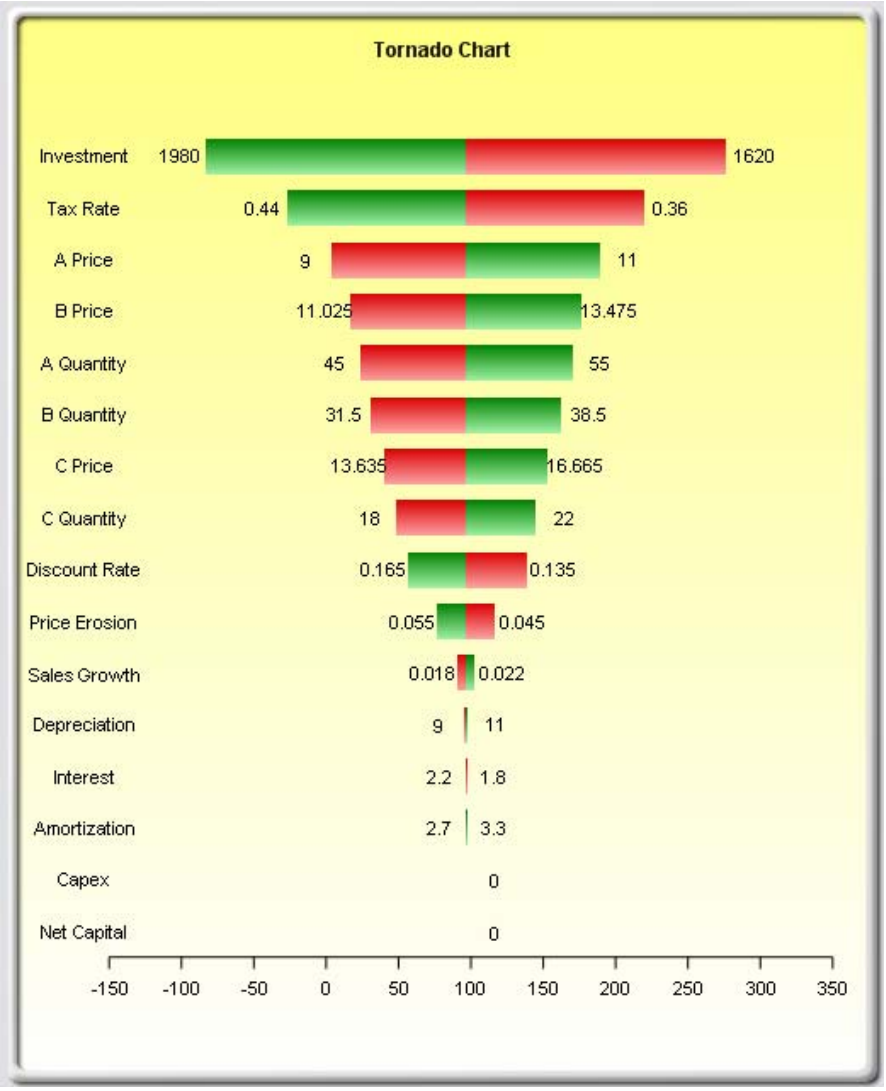


Figure 5.6 – Tornado Chart

Although the tornado chart is easier to read, the spider chart is important for determining if there are any nonlinearities in the model. For instance, Figure 5.7 shows another spider chart where nonlinearities are fairly evident (the lines on the graph are not straight but curved). The model used is Tornado and Sensitivity Charts (Nonlinear), which uses the Black-Scholes option pricing model as an example. Such nonlinearities cannot be ascertained from a tornado chart and may be important information in the model or provide decision makers with important insight into the model's dynamics.

Notes

Figure 5.2 shows the Tornado analysis tool's user interface. Notice that there are a few new enhancements starting in Risk Simulator version 4 and beyond. Here are some tips on running Tornado analysis and details on the new enhancements:

- Tornado analysis should never be run just once. It is meant as a model diagnostic tool, which means that it should ideally be run several times on the same model. For

instance, in a large model, Tornado can be run the first time using all of the default settings and all precedents should be shown (select Show All Variables). The result may be a large report and long (and potentially unsightly) Tornado charts. Nonetheless, this analysis provides a great starting point to determine how many of the precedents are considered critical success factors. For example, the Tornado chart may show that the first 5 variables have high impact on the output, while the remaining 200 variables have little to no impact, in which case, a second Tornado analysis is run showing fewer variables. For example, select the Show Top 10 Variables if the first 5 are critical, thereby creating a nice report and Tornado chart that shows a contrast between the key factors and less critical factors. (You should never show a Tornado chart with only the key variables. You need to show some less critical variables as a contrast to their effects on the output). Finally, the default testing points can be increased from the ±10% of the parameter to some larger value to test for nonlinearities (the Spider chart will show nonlinear lines and Tornado charts will be skewed to one side if the precedent effects are nonlinear).

- *Selecting Use Cell Address* is always a good idea if your model is large, as it allows you to identify the location (worksheet name and cell address) of a precedent cell. If this option is not selected, the software will apply its own fuzzy logic in an attempt to determine the name of each precedent variable (in a large model, the names might sometimes end up being confusing, with repeated variables or the names that are too long, possibly making the Tornado chart unsightly).

- The *Analyze This Worksheet* and *Analyze All Worksheets* options allow you to control whether the precedents should only be part of the current worksheet or include all worksheets in the same workbook. This option comes in handy when you are only attempting to analyze an output based on values in the current sheet versus performing a global search of all linked precedents across multiple worksheets in the same workbook.

- Selecting *Use Global Setting* is useful when you have a large model and wish to test all the precedents at, say, ±50% instead of the default 10%. Instead of having to change each precedent's test values one at a time, you can select this option, change one setting and click somewhere else in the user interface to change the entire list of the precedents. Deselecting this option will allow you the control to change test points one precedent at a time.

- *Ignore Zero or Empty Values* is an option turned on by default where precedent cells with zero or empty values will not be run in the Tornado analysis. This is the typical setting.

- *Highlight Possible Integer Values* is an option that quickly identifies all possible precedent cells that currently have integer inputs. This function is sometimes important if your model uses switches (e.g., functions such as IF a cell is 1. then something happens, and IF a cell has a 0 value, something else happens, or integers such as 1, 2, 3, etc., which you do not wish to test). For instance, ±10% of a flag switch value of 1 will return a test value of 0.9 and 1.1, both of which are irrelevant and incorrect input values in the model, and Excel may interpret the function as an error. This option, when selected, will quickly highlight potential problem areas for Tornado analysis, and then you can determine which precedents to turn on or

off manually, or you can use the Ignore Possible Integer Values function to turn all of them off simultaneously.

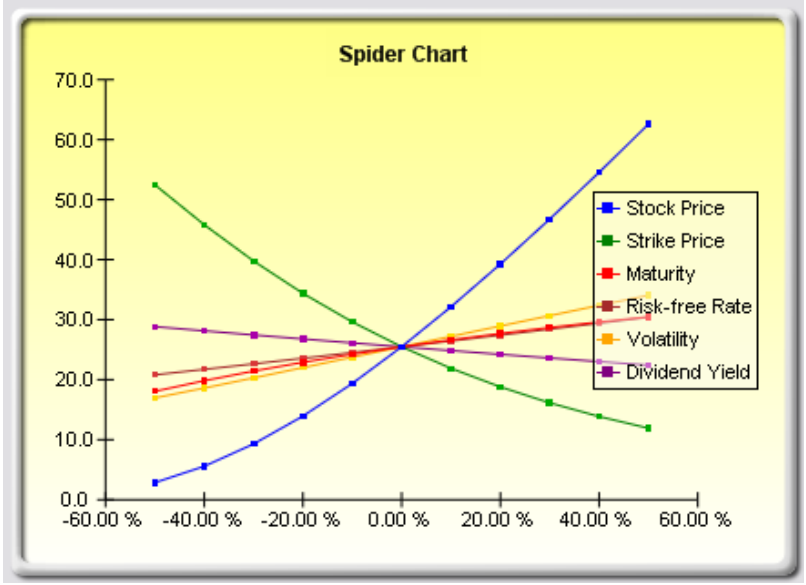


Figure 5.7 – Nonlinear Spider Chart

5.2 Sensitivity Analysis

Theory

While tornado analysis (tornado charts and spider charts) applies static perturbations before a simulation run, sensitivity analysis applies dynamic perturbations created after the simulation run. Tornado and spider charts are the results of static perturbations, meaning that each precedent or assumption variable is perturbed a preset amount one at a time, and the fluctuations in the results are tabulated. In contrast, sensitivity charts are the results of dynamic perturbations in the sense that multiple assumptions are perturbed simultaneously and their interactions in the model and correlations among variables are captured in the fluctuations of the results. Tornado charts, therefore, identify which variables drive the results the most and, hence, are suitable for simulation, whereas sensitivity charts identify the impact to the results when multiple interacting variables are simulated together in the model. This effect is clearly illustrated in Figure 5.8. Notice that the ranking of critical success drivers similar to the tornado chart in the previous examples. However, if correlations are added between the assumptions, a very different picture results, as shown in Figure 5.9. Notice, for instance, that price erosion had little impact on NPV, but when some of the input assumptions are correlated, the interaction that exists between these correlated variables makes price erosion have more impact.

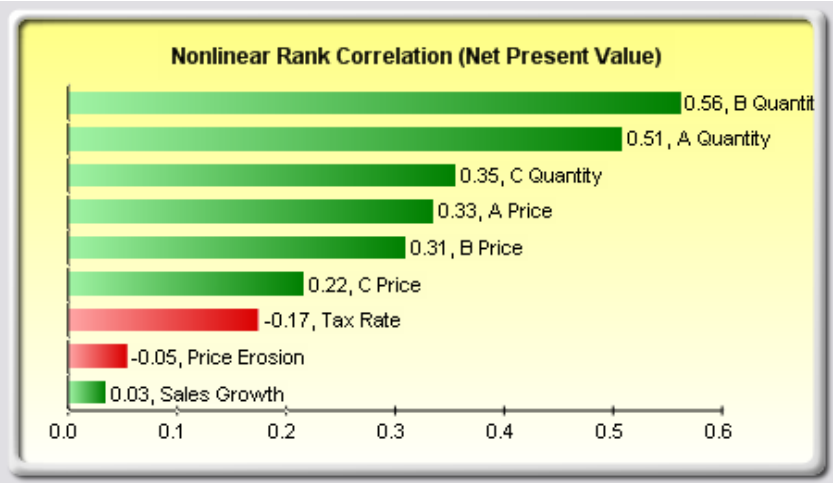


Figure 5.8 – Sensitivity Chart Without Correlations

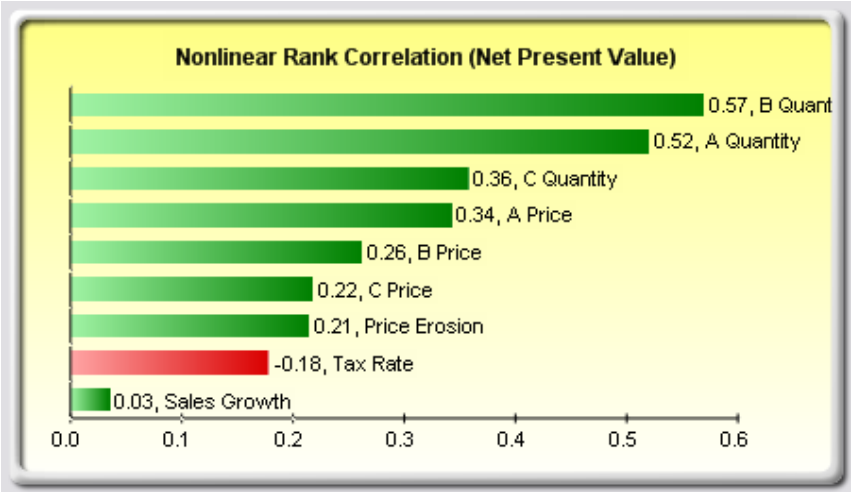


Figure 5.9 – Sensitivity Chart With Correlations

Procedure

- Open or create a model, define assumptions and forecasts, and *run the simulation* (the example here uses the Tornado and Sensitivity Charts (Linear) file).
- Select *Risk Simulator | Tools | Sensitivity Analysis*.
- Select the forecast of choice to analyze and click *OK* (Figure 5.10)

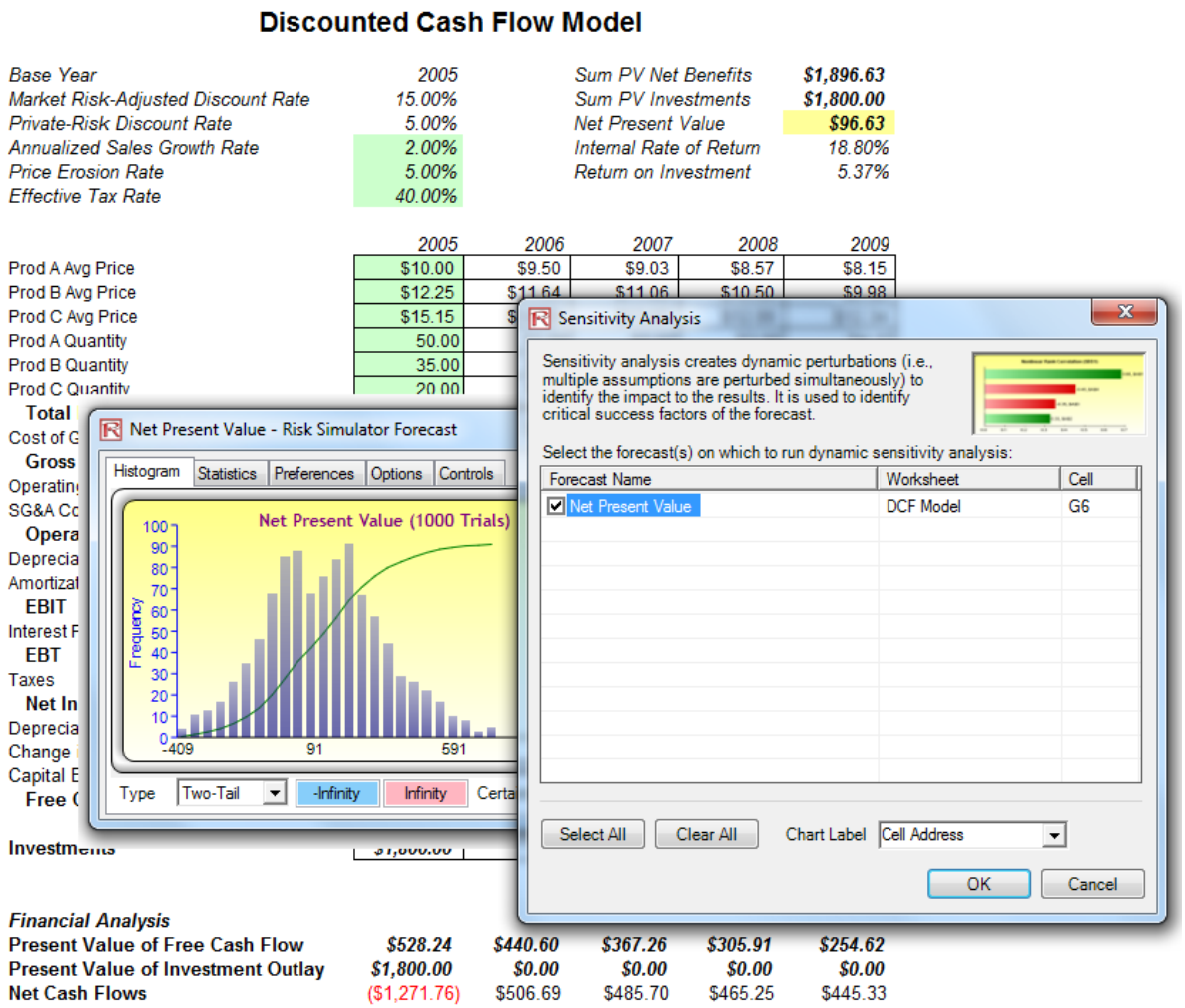


Figure 5.10 – Running Sensitivity Analysis

Results Interpretation

The results of the sensitivity analysis comprise a report and two key charts. The first is a nonlinear rank correlation chart (Figure 5.11) that ranks from highest to lowest the assumption-forecast correlation pairs. These correlations are nonlinear and nonparametric, making them free of any distributional requirements (i.e., an assumption with a Weibull distribution can be compared to another with a beta distribution). The results from this chart are fairly similar to that of the tornado analysis seen previously (of course, without the capital investment value, which we decided was a known value and, hence, was not simulated), with one special exception: Tax rate was relegated to a much lower position in the sensitivity analysis chart (Figure 5.11) as compared to the tornado chart (Figure 5.6). This is because by itself, tax rate will have a significant impact, but once the other variables are interacting in the model, it appears that tax rate has less of a dominant effect (because tax rate has a smaller distribution as historical tax rates tend not to fluctuate too much, and also because tax rate is a straight percentage value of the income before taxes, where other precedent variables have a larger effect on). This example proves that performing sensitivity analysis after a simulation run is important to ascertain if there are any interactions in the model and if the effects of certain variables still hold. The second chart (Figure 5.12) illustrates the percent variation explained. That is, of the fluctuations in the forecast, how much of the variation can be explained by each

of the assumptions after accounting for all the interactions among variables? Notice that the sum of all variations explained is usually close to 100% (there are sometimes other elements that impact the model but that cannot be captured here directly), and if correlations exist, the sum may sometimes exceed 100% (due to the interaction effects that are cumulative).

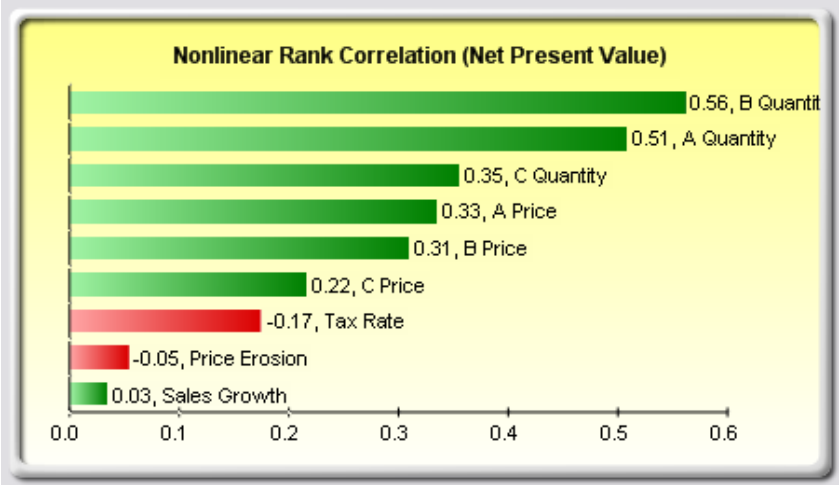


Figure 5.11 – Rank Correlation Chart

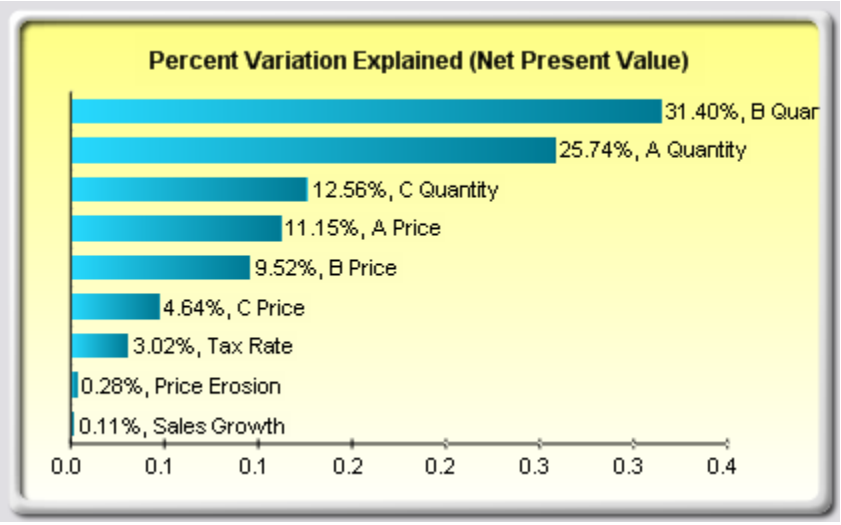


Figure 5.12 – Contribution to Variance Chart

Notes

Tornado analysis is performed before a simulation run, while sensitivity analysis is performed after a simulation run. Spider charts in tornado analysis can consider nonlinearities, while rank correlation charts in sensitivity analysis can account for nonlinear and distributional-free conditions.

5.3 Distributional Fitting: Single Variable and Multiple Variables

Theory

Another powerful simulation tool is distributional fitting. That is, determining which distribution to use for a particular input variable in a model and what the relevant distributional parameters are. If no historical data exist, then the analyst must make assumptions about the variables in question. One approach is to use the Delphi method where a group of experts are tasked with estimating the behavior of each variable. For instance, a group of mechanical engineers can be tasked with evaluating the extreme possibilities of a spring coil's diameter through rigorous experimentation or guesstimates. These values can be used as the variable's input parameters (e.g., uniform distribution with extreme values between 0.5 and 1.2). When testing is not possible (e.g., market share and revenue growth rate), management can still make estimates of potential outcomes and provide the best-case, most-likely case, and worst-case scenarios.

However, if reliable historical data are available, distributional fitting can be accomplished. Assuming that historical patterns hold and that history tends to repeat itself, then historical data can be used to find the best-fitting distribution with their relevant parameters to better define the variables to be simulated. Figures 5.13, 5.14, and 5.15 illustrate a distributional-fitting example. This illustration uses the Data Fitting file in the examples folder.

Procedure

- Open a spreadsheet with existing data for fitting.
- *Select the data* you wish to fit (data should be in a single column with multiple rows).
- Select *Risk Simulator | Tools | Distributional Fitting (Single-Variable).*
- Select the specific distributions you wish to fit to or keep the default where all distributions are selected and click *OK* (Figure 5.13).
- Review the results of the fit, choose the relevant distribution you want, and click *OK* (Figure 5.14).

BIZSTATS Distributional Fitting Methods

Please note that ROV BizStats has some additional methods and algorithms for data fitting including the following methods:

- Distributional Fitting: Akaike Information Criterion (AIC): Rewards goodness of fit but also includes a penalty that is an increasing function of the number of estimated parameters (although AIC penalizes the number of parameters less strongly than other methods).
- Distributional Fitting: Anderson-Darling (AD): When applied to testing if a normal distribution adequately describes a set of data, it is one of the most powerful statistical tools for detecting departures from normality, and is powerful for testing normal tails. However, in non-normal distributions, this test lacks power compared to others.
- Distributional Fitting: Kolmogorov-Smirnov (KS): A nonparametric test for the equality of continuous probability distributions that can be used to compare a sample with a reference probability distribution, making it useful for testing abnormally shaped distributions and non-normal distributions.
- Distributional Fitting: Kuiper's Statistic (K): Related to the KS test making it as sensitive in the tails as at the median and also makes it invariant under cyclic

transformations of the independent variable, making it invaluable when testing for cyclic variations over time. The AD provides equal sensitivity at the tails as the median, but it does not provide the cyclic invariance.

- Distributional Fitting: Schwarz/Bayes Information Criterion (SC/BIC): The SC/BIC introduces a penalty term for the number of parameters in the model with a larger penalty than AIC.

Results Interpretation

The null hypothesis being tested is such that the fitted distribution is the same distribution as the population from which the sample data to be fitted comes. Thus, if the computed p-value is lower than a critical alpha level (typically 0.10 or 0.05), then the distribution is the wrong distribution. Conversely, the higher the p-value, the better the distribution fits the data. Roughly, you can think of p-value as a percentage explained; that is, if the p-value is 0.9727 (Figure 5.14), then setting a normal distribution with a mean of 99.28 and a standard deviation of 10.17 explains about 97.27% of the variation in the data, indicating an especially good fit. Both the results (Figure 5.14) and the report (Figure 5.15) show the test statistic, p-value, theoretical statistics (based on the selected distribution), empirical statistics (based on the raw data), the original data (to maintain a record of the data used), and the assumption complete with the relevant distributional parameters (i.e., if you selected the option to automatically generate assumption and if a simulation profile already exists). The results also rank all the selected distributions and how well they fit the data.

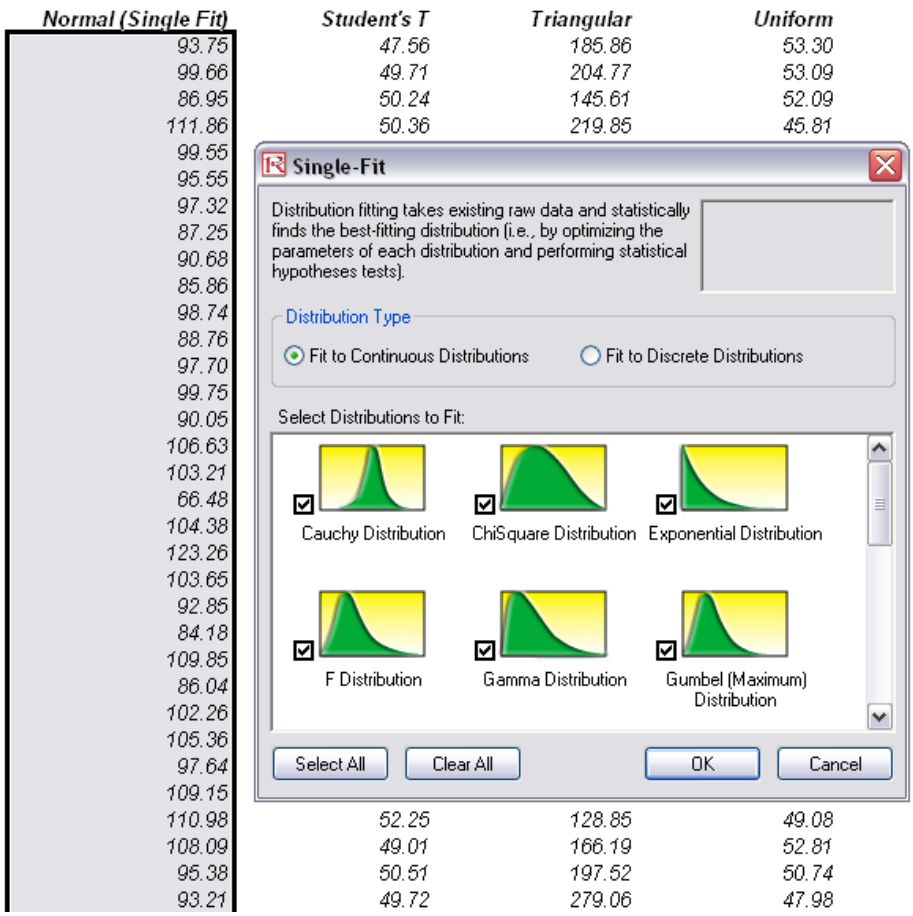
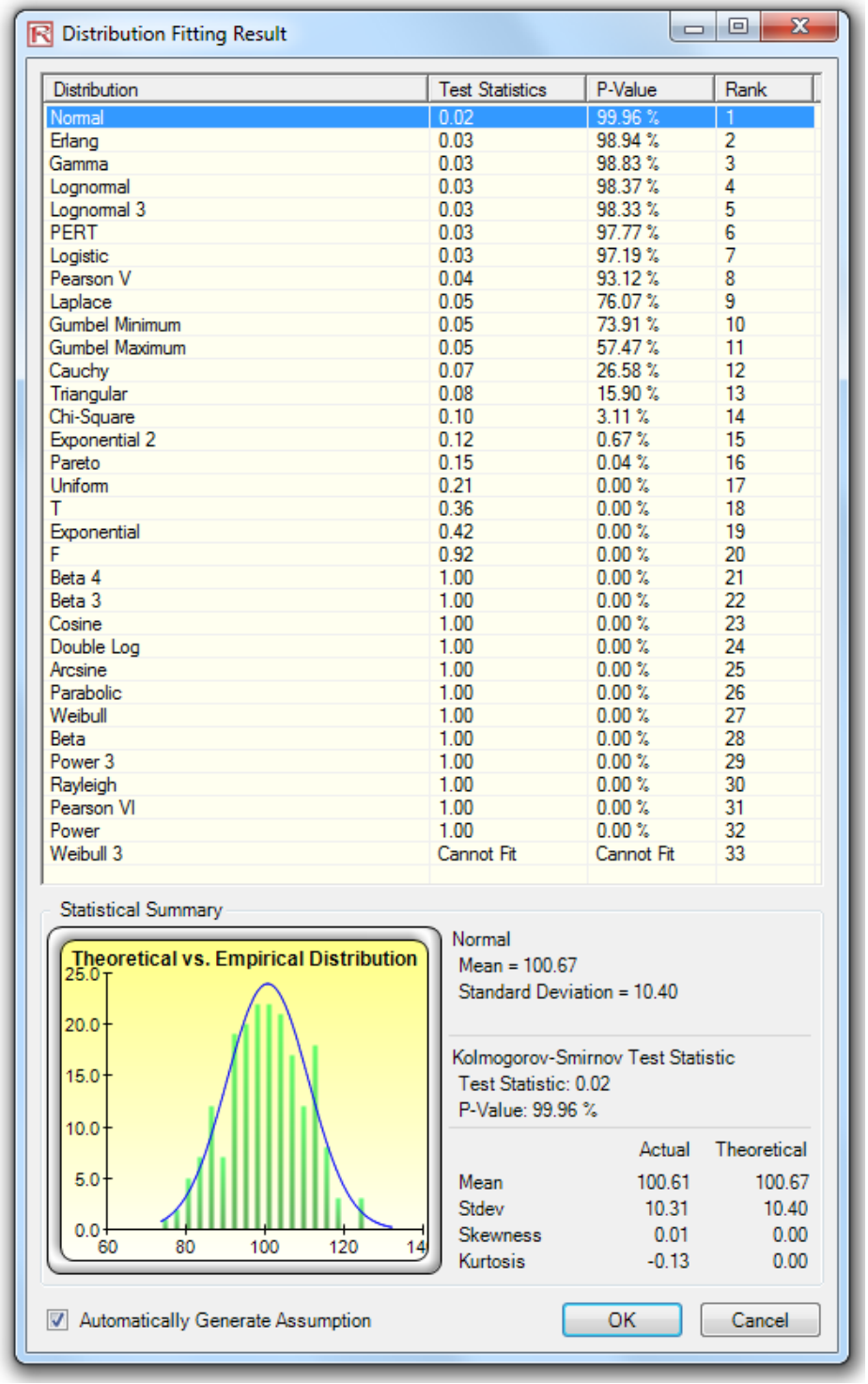


Figure 5.13 – Single Variable Distributional Fitting



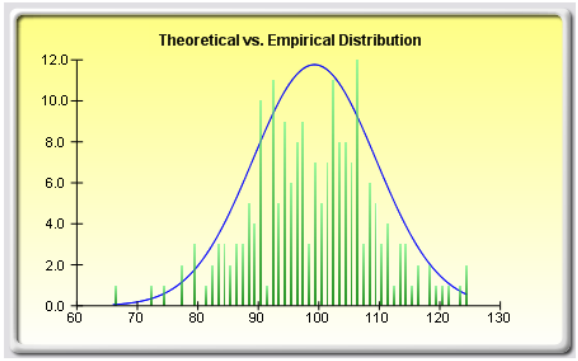
Distribution	Test Statistics	P-Value	Rank
Normal	0.02	99.96 %	1
Erlang	0.03	98.94 %	2
Gamma	0.03	98.83 %	3
Lognormal	0.03	98.37 %	4
Lognormal 3	0.03	98.33 %	5
PERT	0.03	97.77 %	6
Logistic	0.03	97.19 %	7
Pearson V	0.04	93.12 %	8
Laplace	0.05	76.07 %	9
Gumbel Minimum	0.05	73.91 %	10
Gumbel Maximum	0.05	57.47 %	11
Cauchy	0.07	26.58 %	12
Triangular	0.08	15.90 %	13
Chi-Square	0.10	3.11 %	14
Exponential 2	0.12	0.67 %	15
Pareto	0.15	0.04 %	16
Uniform	0.21	0.00 %	17
T	0.36	0.00 %	18
Exponential	0.42	0.00 %	19
F	0.92	0.00 %	20
Beta 4	1.00	0.00 %	21
Beta 3	1.00	0.00 %	22
Cosine	1.00	0.00 %	23
Double Log	1.00	0.00 %	24
Arcsine	1.00	0.00 %	25
Parabolic	1.00	0.00 %	26
Weibull	1.00	0.00 %	27
Beta	1.00	0.00 %	28
Power 3	1.00	0.00 %	29
Rayleigh	1.00	0.00 %	30
Pearson VI	1.00	0.00 %	31
Power	1.00	0.00 %	32
Weibull 3	Cannot Fit	Cannot Fit	33

Figure 5.14 – Distributional Fitting Result

Single Variable Distributional Fitting

Statistical Summary

Fitted Assumption	*99.14*	
Fitted Distribution	***Normal Distribution***	
Mu	*99.28*	
Sigma	*10.17*	
Kolmogorov-Smirnov Statistic	*0.03*	
P-Value for Test Statistic	*0.9727*	
	Actual	*Theoretical*
Mean	*99.14*	*99.28*
Standard Deviation	*10.20*	*10.17*
Skewness	*-0.12*	*0.00*
Excess Kurtosis	*0.10*	*0.00*



Original Fitted Data

93.75	*99.66*	*86.95*	*111.86*	*99.55*	*95.55*	*97.32*	*87.25*	*90.68*	*85.86*	*98.74*	*88.76*	*97.70*
99.75	*90.05*	*106.63*	*103.21*	*66.48*	*104.38*	*123.26*	*103.65*	*92.85*	*84.18*	*109.85*	*86.04*	*102.26*
105.36	*97.64*	*109.15*	*110.98*	*108.09*	*95.38*	*93.21*	*83.86*	*100.17*	*110.17*	*103.72*	*120.52*	*95.09*
115.18	*83.64*	*90.23*	*92.44*	*92.37*	*92.70*	*110.81*	*72.67*	*104.23*	*96.47*	*121.15*	*94.92*	*77.26*
103.45	*96.75*	*93.91*	*101.91*	*124.14*	*90.95*	*107.13*	*92.02*	*96.43*	*96.35*	*88.30*	*108.48*	*113.50*
101.40	*104.72*	*102.43*	*113.59*	*124.15*	*109.24*	*105.34*	*104.57*	*97.83*	*94.39*	*116.19*	*84.66*	*101.17*
106.13	*107.17*	*95.83*	*106.67*	*92.42*	*79.64*	*94.15*	*106.00*	*113.45*	*92.63*	*94.51*	*93.05*	*96.19*
100.85	*83.34*	*111.82*	*118.12*	*87.17*	*103.66*	*106.93*	*82.45*	*102.74*	*86.82*	*106.68*	*89.61*	*94.56*
101.34	*91.32*	*102.02*	*82.51*	*104.46*	*84.72*	*105.05*	*108.40*	*106.59*	*109.43*	*92.49*	*94.52*	*94.00*
105.92	*88.13*	*96.41*	*101.45*	*79.93*	*89.68*	*102.91*	*114.95*	*92.58*	*94.05*	*107.90*	*111.05*	*90.58*
97.09	*105.44*	*94.95*	*102.55*	*77.41*	*108.53*	*90.54*	*100.41*	*106.83*	*99.63*	*79.72*	*89.32*	*116.30*
98.27	*101.73*	*90.84*	*74.45*	*102.24*	*103.34*	*96.51*	*114.55*	*93.94*	*106.29*	*102.95*	*112.73*	*98.09*
108.20	*105.80*	*106.48*	*102.88*	*104.93*	*103.00*	*99.10*	*108.52*	*101.31*	*88.17*	*90.62*	*96.53*	*106.03*
109.12	*104.23*	*90.34*	*95.12*	*102.03*	*100.00*	*118.17*	*99.06*	*81.89*	*104.29*	*92.68*	*114.89*	*102.49*
119.21	*106.20*	*88.26*	*92.45*	*105.15*	*103.79*	*100.84*	*95.19*	*85.10*	*97.25*	*87.65*	*97.58*	*111.44*
99.52	*89.83*	*97.86*	*90.96*	*97.14*								

Figure 5.15 – Distributional Fitting Report

For fitting multiple variables, the process is fairly similar to fitting individual variables. However, the data should be arranged in columns (i.e., each variable is arranged as a column) and all the variables are fitted one at a time.

Procedure

- Open a spreadsheet with existing data for fitting.
- *Select the data* you wish to fit (data should be in a multiple columns with multiple rows).
- Select *Risk Simulator | Tools | Distributional Fitting (Multi-Variable).*
- Review the data, choose the relevant types of distribution you want and click *OK*.

Notes

Notice that the statistical ranking methods used in the distributional fitting routines are the chi-square test and Kolmogorov-Smirnov test. The former is used to test discrete distributions and the latter, continuous distributions. Briefly, a hypothesis test coupled with an internal optimization routine is used to find the best-fitting parameters on each distribution tested, and the results are ranked from the best fit to the worst fit.

5.4 Bootstrap Simulation

Theory

Bootstrap simulation is a simple technique that estimates the reliability or accuracy of forecast statistics or other sample raw data. Essentially, bootstrap simulation is used in hypothesis testing. Classical methods used in the past relied on mathematical formulas to describe the accuracy of sample statistics. These methods assume that the distribution of a sample statistic approaches a normal distribution, making the calculation of the statistic's standard error or confidence interval relatively easy. However, when a statistic's sampling distribution is not normally distributed or easily found, these classical methods are difficult to use or are invalid. In contrast, bootstrapping analyzes sample statistics empirically by repeatedly sampling the data and creating distributions of the different statistics from each sampling.

Procedure

- *Run a simulation*.
- Select *Risk Simulator | Tools | Nonparametric Bootstrap*.
- Select only one *forecast* to bootstrap, select the *statistic(s)* to bootstrap, enter the number of bootstrap trials, and click *OK* (Figure 5.16).

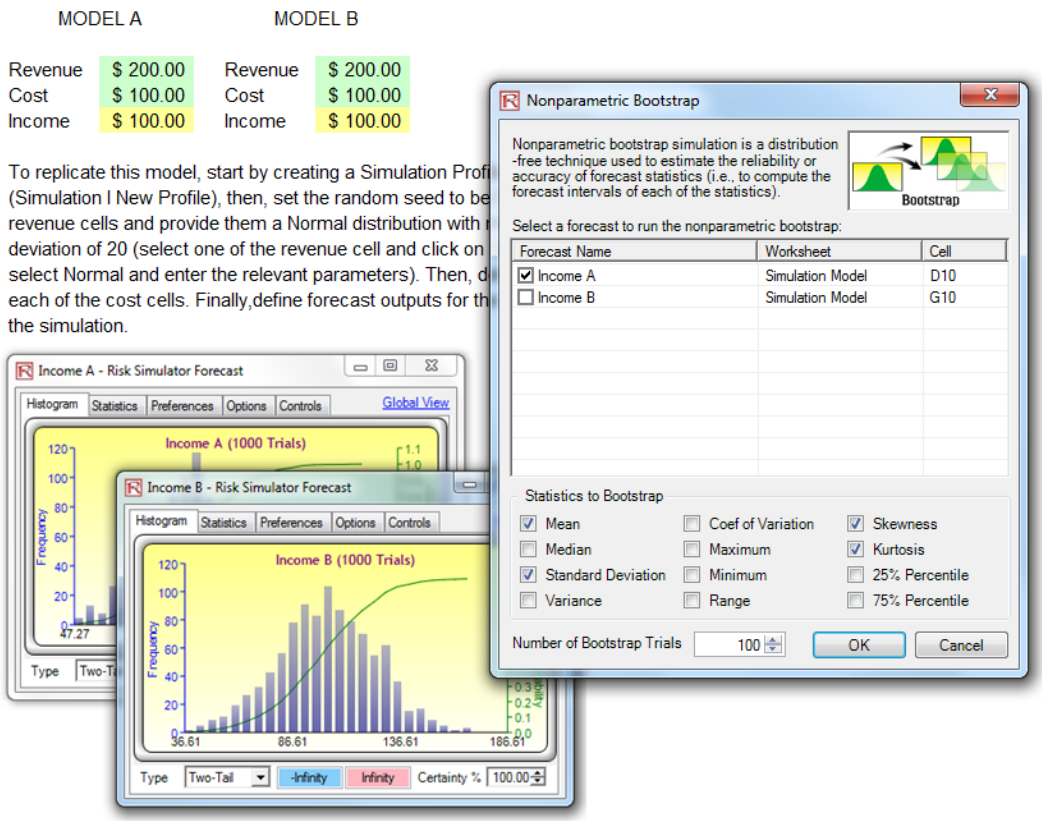


Figure 5.16 – Nonparametric Bootstrap Simulation

Results Interpretation

In essence, nonparametric bootstrap simulation can be thought of as simulation based on a simulation. Thus, after running a simulation, the resulting statistics are displayed, but the accuracy of such statistics and their statistical significance are sometimes in question. For instance, if a simulation run's skewness statistic is –0.10, is this distribution truly negatively skewed or is the slight negative value attributable to random chance? What about –0.15, –0.20, and so forth? That is, how far is far enough such that this distribution is considered to be

negatively skewed? The same question can be applied to all the other statistics. Is one distribution statistically identical to another distribution with regard to some computed statistics or are they significantly different? Suppose for instance, the 90% confidence for the skewness statistic is between –0.0189 and 0.0952, such that the value 0 falls within this confidence, indicating that on a 90% confidence, the skewness of this forecast is not statistically significantly different from 0, or that this distribution can be considered as symmetrical and not skewed. Conversely, if the value 0 falls outside of this confidence, then the opposite is true, and the distribution is skewed (positively skewed if the forecast statistic is positive, and negatively skewed if the forecast statistic is negative). Figure 5.17 illustrates some sample bootstrap results.

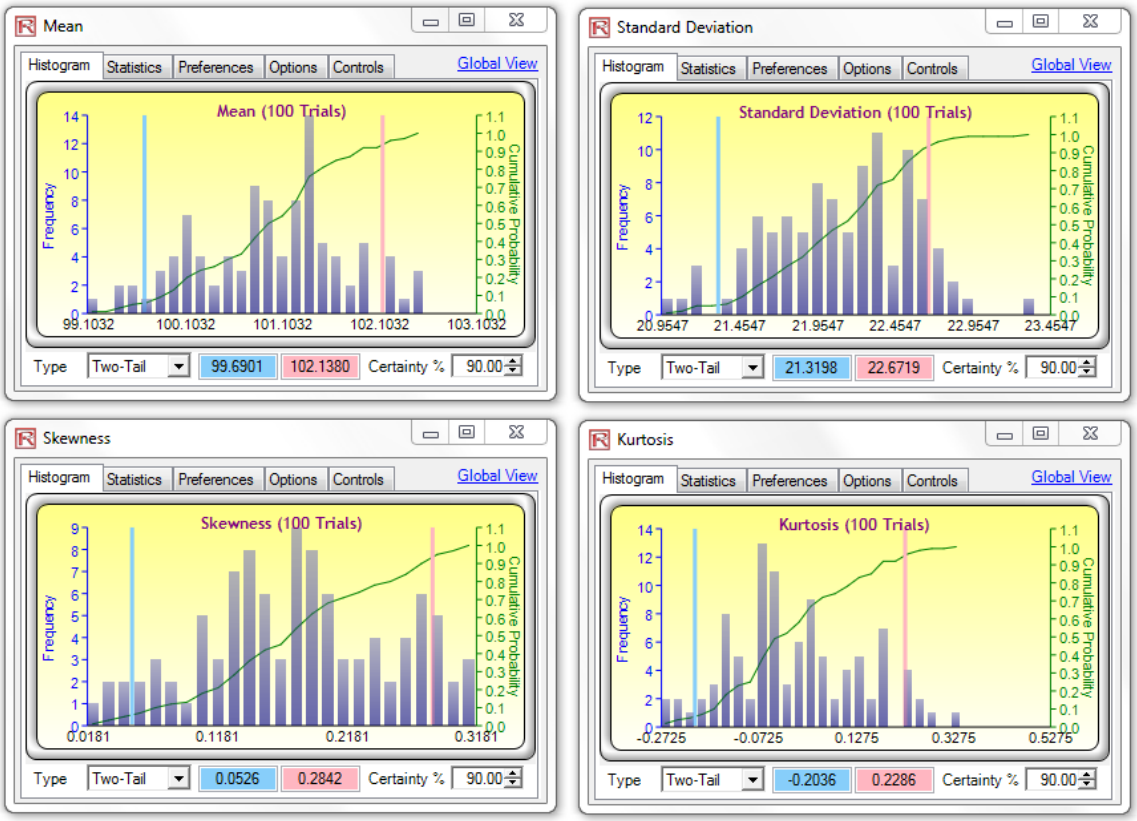


Figure 5.17 – Bootstrap Simulation Results

Notes

The term bootstrap comes from the saying, "to pull oneself up by one's own bootstraps," and is applicable because this method uses the distribution of statistics themselves to analyze the statistics' accuracy. Nonparametric simulation is simply randomly picking golf balls from a large basket with replacement where each golf ball is based on a historical data point. Suppose there are 365 golf balls in the basket (representing 365 historical data points). Imagine that the value of each golf ball picked at random is written on a large whiteboard. The results of the 365 balls picked with replacement are written in the first column of the board with 365 rows of numbers. Relevant statistics (e.g., mean, median, standard deviation, etc.) are calculated on these 365 rows. The process is then repeated, say, five thousand times. The whiteboard will now be filled with 365 rows and 5,000 columns. Hence, 5,000 sets of statistics (i.e., there will be 5,000 means, 5,000 medians, 5,000 standard deviations, etc.) are tabulated and their distributions shown. The relevant statistics of the statistics are then tabulated, where from these results one can ascertain

how confident the simulated statistics are. In other words, in a simple 10,000-trial simulation, say the resulting forecast average is found to be $5.00. How certain is the analyst of the results? Bootstrapping allows the user to ascertain the confidence interval of the calculated mean statistic, indicating the distribution of the statistics. Finally, bootstrap results are important because according to the Law of Large Numbers and the Central Limit Theorem in statistics, the mean of the sample means is an unbiased estimator and approaches the true population mean when the sample size increases.

5.5 Hypothesis Testing

Theory

A hypothesis test is performed when testing the means and variances of two distributions to determine if they are statistically identical or statistically different from one another; that is, whether the differences are based on random chance or if they are, in fact, statistically significant.

Procedure

- *Run a simulation*.
- Select *Risk Simulator | Tools | Hypothesis Testing*.
- Select only *two forecasts* to test at a time, select the type of hypothesis test you wish to run, and click *OK* (Figure 5.18).

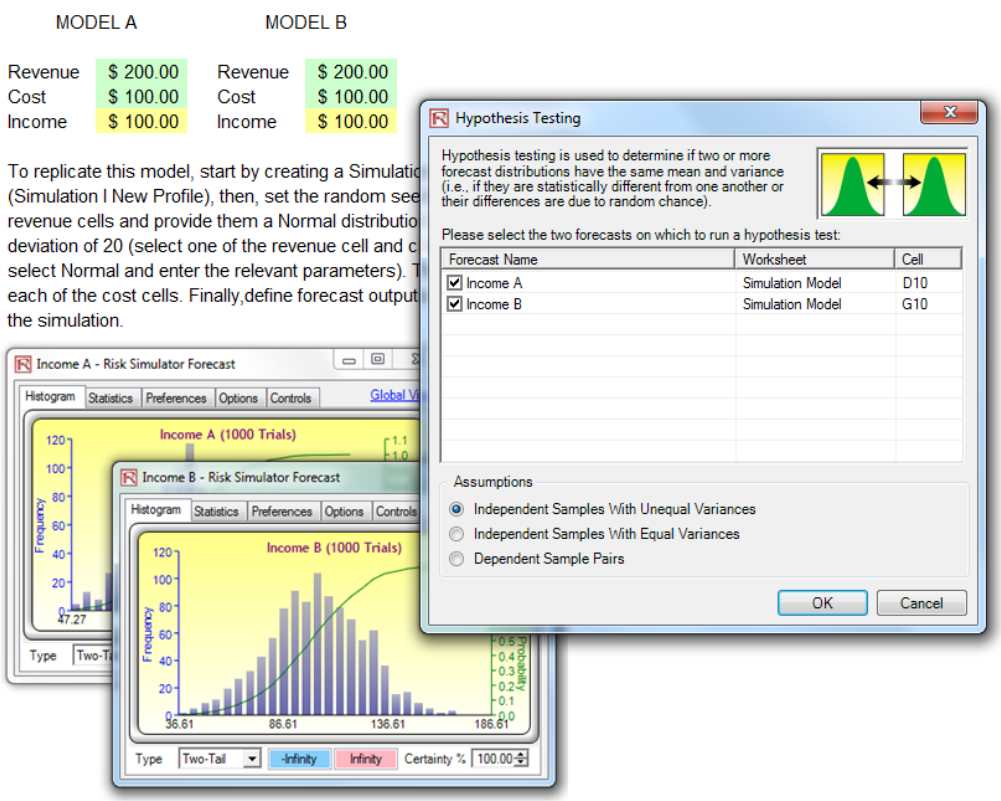


Figure 5.18 – Hypothesis Testing

Results Interpretation

A two-tailed hypothesis test is performed on the null hypothesis (H0) such that the two variables' population means are statistically identical to one another. The alternative hypothesis (Ha) is such that the population means are statistically different from one another. If the calculated p-values are less than or equal to 0.01, 0.05, or 0.10, this means that the null hypothesis is rejected, which implies that the forecast means are statistically significantly different at the 1%, 5%, and 10% significance levels. If the null hypothesis is not rejected when the p-values are high, the means of the two forecast distributions are statistically similar to one another. The same analysis is performed on variances of two forecasts at a time using the pairwise F-test. If the p-values are small, then the variances (and standard deviations) are statistically different from one another; otherwise, for large p-values, the variances are statistically identical to one another.

Hypothesis Test on the Means and Variances of Two Forecasts

Statistical Summary

A hypothesis test is performed when testing the means and variances of two distributions to determine if they are statistically identical or statistically different from one another. That is, to see if the differences between two means and two variances that occur are based on random chance or they are in fact different from one another. The two-variable t-test with unequal variances (the population variance of forecast 1 is expected to be different from the population variance of forecast 2) is appropriate when the forecast distributions are from different populations (e.g., data collected from two different geographical locations, two different operating business units, and so forth). The two-variable t-test with equal variances (the population variance of forecast 1 is expected to be equal to the population variance of forecast 2) is appropriate when the forecast distributions are from similar populations (e.g., data collected from two different engine designs with similar specifications, and so forth). The paired dependent two-variable t-test is appropriate when the forecast distributions are from similar populations (e.g., data collected from the same group of customers but on different occasions, and so forth).

A two-tailed hypothesis test is performed on the null hypothesis Ho such that the two variables' population means are statistically identical to one another. The alternative hypothesis is that the population means are statistically different from one another. If the calculated p-values are less than or equal to 0.01, 0.05, or 0.10, this means that the hypothesis is rejected, which implies that the forecast means are statistically significantly different at the 1%, 5% and 10% significance levels. If the null hypothesis is not rejected when the p-values are high, the means of the two forecast distributions are statistically similar to one another. The same analysis is performed on variances of two forecasts at a time using the pairwise F-Test. If the p-values are small, then the variances (and standard deviations) are statistically different from one another, otherwise, for large p-values, the variances are statistically identical to one another.

Result

Hypothesis Test Assumption:	*Unequal Variances:*
Computed t-statistic:	*-0.32947*
P-value for t-statistic:	*0.74181*
Computed F-statistic:	*1.026723*
P-value for F-statistic:	*0.351212*

Figure 5.19 – Hypothesis Testing Results

Notes

The two-variable t-test with unequal variances (the population variance of forecast 1 is expected to be different from the population variance of forecast 2) is appropriate when the forecast distributions are from different populations (e.g., data collected from two different geographical locations or two different operating business units). The two-variable t-test with equal variances (the population variance of forecast 1 is expected to be equal to the population variance of forecast 2) is appropriate when the forecast distributions are from similar populations (e.g., data collected from two different engine designs with similar specifications). The paired dependent two-variable t-test is appropriate when the forecast distributions are from the exact same population (e.g., data collected from the same group of customers but on different occasions).

5.6 Data Extraction and Saving Simulation Results

A simulation's raw data can be very easily extracted using Risk Simulator's Data Extraction routine. Both assumptions and forecasts can be extracted, but a simulation must first be run. The extracted data can then be used for a variety of other analysis.

Procedure

- Open or create a model, define assumptions and forecasts, and *run the simulation*.
- Select *Risk Simulator | Tools | Data Extraction*.
- Select the *assumptions* and/or *forecasts* you wish to extract the data from and click *OK*.

The data can be extracted to various formats:

- Raw data in a new worksheet where the simulated values (both assumptions and forecasts) can then be saved or further analyzed as required
- Flat text file where the data can be exported into other data analysis software
- Risk Simulator file where the results (both assumptions and forecasts) can be retrieved at a later time by selecting *Risk Simulator | Tools | Data Open/Import*

The third option is the most popular selection, that is, to save the simulated results as a *.risksim file where the results can be retrieved later and a simulation does not have to be rerun each time. Figure 5.20 shows the dialog box for extracting or exporting and saving the simulation results.

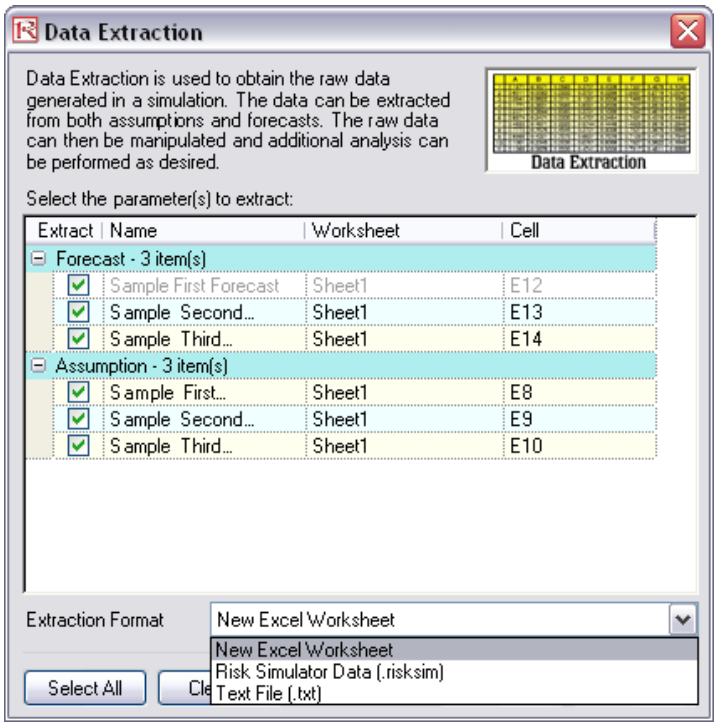


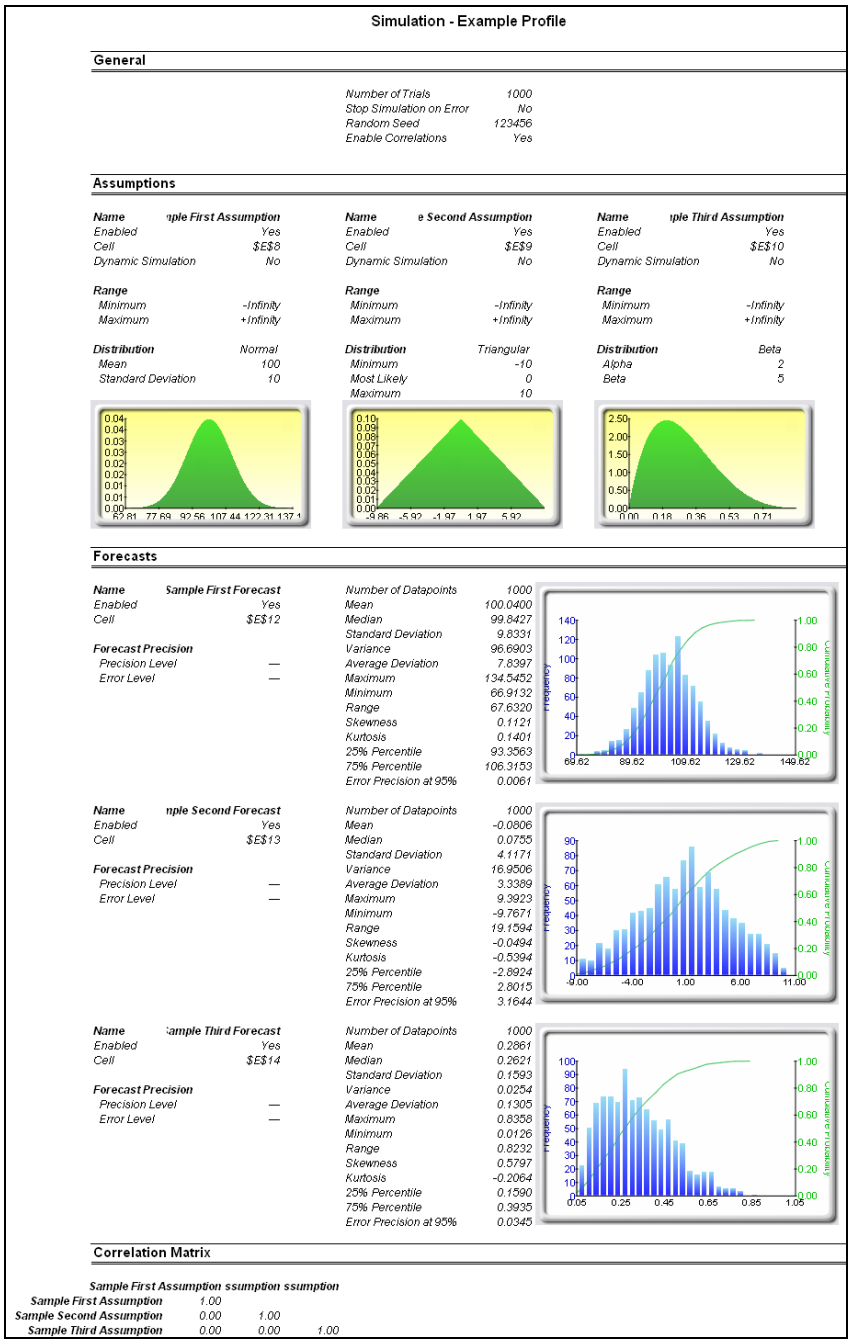
Figure 5.20 – Sample Simulation Report

5.7 Create Report

After a simulation is run, you can generate a report of the assumptions and forecasts used in the simulation run, as well as the results obtained during the simulation run.

Procedure

- Open or create a model, define assumptions and forecasts, and *run the simulation*.
- Select *Risk Simulator | Create Report* (Figure 5.21).



Simulation - Example Profile

General

Number of Trials	1000
Stop Simulation on Error	No
Random Seed	123456
Enable Correlations	Yes

Assumptions

Name	ple First Assumption	e Second Assumption	ple Third Assumption
Enabled	Yes	Yes	Yes
Cell	E8	E9	E10
Dynamic Simulation	No	No	No
Range			
Minimum	-Infinity	-Infinity	-Infinity
Maximum	+Infinity	+Infinity	+Infinity
Distribution	Normal	Triangular	Beta
Mean	100	Minimum -10	Alpha 2
Standard Deviation	10	Most Likely 0	Beta 5
		Maximum 10	

Forecasts

Name	Sample First Forecast	Sample Second Forecast	Sample Third Forecast
Enabled	Yes	Yes	Yes
Cell	E12	E13	E14
Forecast Precision			
Precision Level	—	—	—
Error Level	—	—	—
Number of Datapoints	1000	1000	1000
Mean	100.0400	-0.0806	0.2861
Median	99.8427	0.0755	0.2621
Standard Deviation	9.8331	4.1171	0.1593
Variance	96.6903	16.9506	0.0254
Average Deviation	7.8397	3.3389	0.1305
Maximum	134.5452	9.3923	0.8358
Minimum	66.9132	-9.7671	0.0126
Range	67.6320	19.1594	0.8232
Skewness	0.1121	-0.0494	0.5797
Kurtosis	0.1401	-0.5394	-0.2064
25% Percentile	93.3563	-2.8924	0.1590
75% Percentile	106.3153	2.8015	0.3935
Error Precision at 95%	0.0061	3.1644	0.0345

Correlation Matrix

	Sample First Assumption	ssumption	ssumption
Sample First Assumption	1.00		
Sample Second Assumption	0.00	1.00	
Sample Third Assumption	0.00	0.00	1.00

Figure 5.21 – Sample Simulation Report

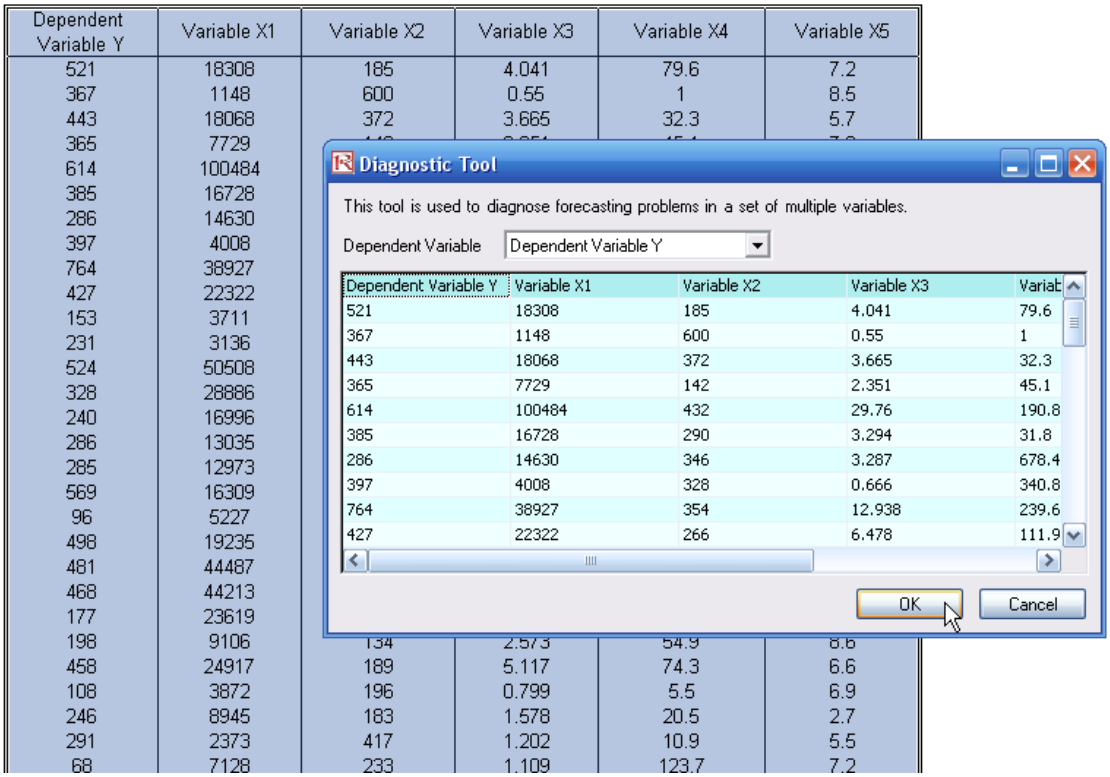
5.8 Regression and Forecasting Diagnostic Tool

The regression and forecasting Diagnostic tool in Risk Simulator is an advanced analytical tool used to determine the econometric properties of your data. The diagnostics include checking the data for heteroskedasticity, nonlinearity, outliers, specification errors, micronumerosity, stationarity and stochastic properties, normality and sphericity of the errors, and multicollinearity. Each test is described in more detail in its respective report in the model.

Procedure

- Open the example model (*Risk Simulator | Examples | Regression Diagnostics*), go to the Time-Series Data worksheet, and select the data, including the variable names (cells *C5:H55*).
- Click on *Risk Simulator | Tools | Diagnostic Tool*.
- Check the data and select from the *Dependent Variable Y* drop-down menu. Click *OK* when finished (Figure 5.22).

Multiple Regression Analysis Data Set



Dependent Variable Y	Variable X1	Variable X2	Variable X3	Variable X4	Variable X5
521	18308	185	4.041	79.6	7.2
367	1148	600	0.55	1	8.5
443	18068	372	3.665	32.3	5.7
365	7729				
614	100484				
385	16728				
286	14630				
397	4008				
764	38927				
427	22322				
153	3711				
231	3136				
524	50508				
328	28886				
240	16996				
286	13035				
285	12973				
569	16309				
96	5227				
498	19235				
481	44487				
468	44213				
177	23619				
198	9106	134	2.573	54.9	6.6
458	24917	189	5.117	74.3	6.6
108	3872	196	0.799	5.5	6.9
246	8945	183	1.578	20.5	2.7
291	2373	417	1.202	10.9	5.5
68	7128	233	1.109	123.7	7.2

Figure 5.22 – Running the Data Diagnostic Tool

A common violation in forecasting and regression analysis is heteroskedasticity, that is, the variance of the errors increases over time (see Figure 5.23 for test results using the Diagnostic tool). Visually, the width of the vertical data fluctuations increases, or fans out, over time, and, typically, the coefficient of determination (R-squared coefficient) drops significantly when heteroskedasticity exists. If the variance of the dependent variable is not constant, then the

error's variance will not be constant. Unless the heteroskedasticity of the dependent variable is pronounced, its effect will not be severe: The least-squares estimates will still be unbiased, and the estimates of the slope and intercept will either be normally distributed if the errors are normally distributed, or at least normally distributed asymptotically (as the number of data points becomes large) if the errors are not normally distributed. The estimate for the variance of the slope and overall variance will be inaccurate, but the inaccuracy is not likely to be substantial if the independent-variable values are symmetric about their mean.

If the number of data points is small (micronumerosity), it may be difficult to detect assumption violations. With small sample sizes, assumption violations such as non-normality or heteroskedasticity of variances are difficult to detect even when they are present. With a small number of data points, linear regression offers less protection against violation of assumptions. With few data points, it may be hard to determine how well the fitted line matches the data, or whether a nonlinear function would be more appropriate. Even if none of the test assumptions are violated, a linear regression on a small number of data points may not have sufficient power to detect a significant difference between the slope and zero, even if the slope is nonzero. The power depends on the residual error, the observed variation in the independent variable, the selected significance alpha level of the test, and the number of data points. Power decreases as the residual variance increases, decreases as the significance level is decreased (i.e., as the test is made more stringent), increases as the variation in observed independent variable increases, and increases as the number of data points increases.

Values may not be identically distributed because of the presence of outliers which are anomalous values in the data. Outliers may have a strong influence over the fitted slope and intercept, giving a poor fit to the bulk of the data points. Outliers tend to increase the estimate of residual variance, lowering the chance of rejecting the null hypothesis (that is, creating higher prediction errors). They may be due to recording errors, which may be correctable, or they may be due to the dependent-variable values not all being sampled from the same population. Apparent outliers may also be due to the dependent-variable values being from the same, but non-normal, population. However, a point may be an unusual value in either an independent or dependent variable without necessarily being an outlier in the scatter plot. In regression analysis, the fitted line can be highly sensitive to outliers. In other words, least squares regression is not resistant to outliers, thus, neither is the fitted-slope estimate. A point vertically removed from the other points can cause the fitted line to pass close to it, instead of following the general linear trend of the rest of the data, especially if the point is relatively far horizontally from the center of the data.

However, great care should be taken when deciding if the outliers should be removed. Although in most cases when outliers are removed, the regression results look better, a priori justification must first exist. For instance, if one is regressing the performance of a particular firm's stock returns, outliers caused by downturns in the stock market should be included; these are not truly outliers as they are inevitabilities in the business cycle. Forgoing these outliers and using the regression equation to forecast one's retirement fund based on the firm's stocks will yield incorrect results at best. In contrast, suppose the outliers are caused by a single nonrecurring business condition (e.g., merger and acquisition) and such business structural changes are not forecast to recur. These outliers, then, should be removed and the data cleansed prior to running a regression analysis. The analysis here only identifies outliers and it is up to the user to determine if they should remain or be excluded.

Sometimes, a nonlinear relationship between the dependent and independent variables is more appropriate than a linear relationship. In such cases, running a linear regression will not be optimal. If the linear model is not the correct form, then the slope and intercept estimates and

the fitted values from the linear regression will be biased, and the fitted slope and intercept estimates will not be meaningful. Over a restricted range of independent or dependent variables, nonlinear models may be well approximated by linear models (this is, in fact, the basis of linear interpolation), but for accurate prediction, a model appropriate to the data should be selected. A nonlinear transformation should first be applied to the data before running a regression. One simple approach is to take the natural logarithm of the independent variable (other approaches include taking the square root or raising the independent variable to the second or third power) and run a regression or forecast using the nonlinearly transformed data.

Diagnostic Results

	Heteroskedasticity		Micronumerosity	Outliers			Nonlinearity	
Variable	W-Test p-value	Hypothesis Test result	Approximation result	Natural Lower Bound	Natural Upper Bound	Number of Potential Outliers	Nonlinear Test p-value	Hypothesis Test result
Y			no problems	-7.86	671.70	2		
Variable X1	0.2543	Homoskedastic	no problems	-21377.95	64713.03	3	0.2458	linear
Variable X2	0.3371	Homoskedastic	no problems	77.47	445.93	2	0.0335	nonlinear
Variable X3	0.3649	Homoskedastic	no problems	-5.77	15.69	3	0.0305	nonlinear
Variable X4	0.3066	Homoskedastic	no problems	-295.96	628.21	4	0.9298	linear
Variable X5	0.2495	Homoskedastic	no problems	3.35	9.38	3	0.2727	linear

Figure 5.23 – Results from Tests of Outliers, Heteroskedasticity, Micronumerosity, and Nonlinearity

Another typical issue when forecasting time-series data is whether the independent-variable values are truly independent of each other or are actually dependent. Dependent variable values collected over a time series may be autocorrelated. For serially correlated dependent-variable values, the estimates of the slope and intercept will be unbiased, but the estimates of their forecast and variances will not be reliable and, hence, the validity of certain statistical goodness-of-fit tests will be flawed. For instance, interest rates, inflation rates, sales, revenues, and many other time-series data are typically autocorrelated, where the value in the current period is related to the value in a previous period, and so forth (clearly, the inflation rate in March is related to February's level, which, in turn, is related to January's level, etc.). Ignoring such blatant relationships will yield biased and less accurate forecasts. In such events, an autocorrelated regression model, or an ARIMA model, may be better suited (*Risk Simulator | Forecasting | ARIMA*). Finally, the autocorrelation functions of a series that is nonstationary tend to decay slowly (see the nonstationary report in the model).

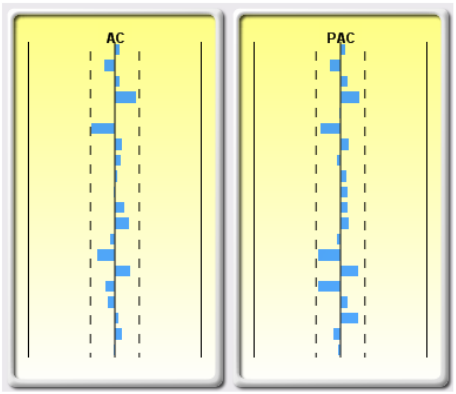
If autocorrelation AC(1) is nonzero, it means that the series is first-order serially correlated. If AC(k) dies off more or less geometrically with increasing lag, it implies that the series follows a low-order autoregressive process. If AC(k) drops to zero after a small number of lags, it implies that the series follows a low-order moving-average process. Partial correlation PAC(k) measures the correlation of values that are k periods apart after removing the correlation from the intervening lags. If the pattern of autocorrelation can be captured by an autoregression of order less than k, then the partial autocorrelation at lag k will be close to zero. Ljung-Box Q-statistics and their p-values at lag k have the null hypothesis that there is no autocorrelation up to order k. The dotted lines in the plots of the autocorrelations are the approximate two standard error bounds. If the autocorrelation is within these bounds, it is not significantly different from zero at the 5% significance level.

Autocorrelation measures the relationship to the past of the dependent Y variable to itself. Distributive lags, in contrast, are time-lag relationships between the dependent Y variable and different independent X variables. For instance, the movement and direction of mortgage rates tend to follow the federal funds rate but at a time lag (typically 1 to 3 months). Sometimes, time lags follow cycles and seasonality (e.g., ice cream sales tend to peak during the summer months

and are, hence, related to last summer's sales, 12 months in the past). The distributive lag analysis (Figure 5.24) shows how the dependent variable is related to each of the independent variables at various time lags, when all lags are considered simultaneously, to determine which time lags are statistically significant and should be considered.

Autocorrelation

Time Lag	AC	PAC	Lower Bound	Upper Bound	Q-Stat	Prob
1	0.0580	0.0580	-0.2828	0.2828	0.1786	0.6726
2	-0.1213	-0.1251	-0.2828	0.2828	0.9754	0.6140
3	0.0590	0.0756	-0.2828	0.2828	1.1679	0.7607
4	0.2423	0.2232	-0.2828	0.2828	4.4865	0.3442
5	0.0067	-0.0078	-0.2828	0.2828	4.4890	0.4814
6	-0.2654	-0.2345	-0.2828	0.2828	8.6516	0.1941
7	0.0814	0.0939	-0.2828	0.2828	9.0524	0.2489
8	0.0634	-0.0442	-0.2828	0.2828	9.3012	0.3175
9	0.0204	0.0673	-0.2828	0.2828	9.3276	0.4076
10	-0.0190	0.0865	-0.2828	0.2828	9.3512	0.4991
11	0.1035	0.0790	-0.2828	0.2828	10.0648	0.5246
12	0.1658	0.0978	-0.2828	0.2828	11.9466	0.4500
13	-0.0524	-0.0430	-0.2828	0.2828	12.1394	0.5162
14	-0.2050	-0.2523	-0.2828	0.2828	15.1738	0.3664
15	0.1782	0.2089	-0.2828	0.2828	17.5315	0.2881
16	-0.1022	-0.2591	-0.2828	0.2828	18.3296	0.3050
17	-0.0861	0.0808	-0.2828	0.2828	18.9141	0.3335
18	0.0418	0.1987	-0.2828	0.2828	19.0559	0.3884
19	0.0869	-0.0821	-0.2828	0.2828	19.6894	0.4135
20	-0.0091	-0.0269	-0.2828	0.2828	19.6966	0.4770



Distributive Lags

P-Values of Distributive Lag Periods of Each Independent Variable

Variable	1	2	3	4	5	6	7	8	9	10	11	12
X1	0.8467	0.2045	0.3336	0.9105	0.9757	0.1020	0.9205	0.1267	0.5431	0.9110	0.7495	0.4016
X2	0.6077	0.9900	0.8422	0.2851	**0.0638**	**0.0032**	0.8007	0.1551	0.4823	0.1126	**0.0519**	0.4383
X3	0.7394	0.2396	0.2741	0.8372	0.9808	**0.0464**	0.8355	**0.0545**	0.6828	0.7354	0.5093	0.3500
X4	**0.0061**	0.6739	0.7932	0.7719	0.6748	0.8627	0.5586	0.9046	0.5726	0.6304	0.4812	0.5707
X5	0.1591	0.2032	0.4123	0.5599	0.6416	0.3447	0.9190	0.9740	0.5185	0.2856	0.1489	0.7794

Figure 5.24 – Autocorrelation and Distributive Lag Results

Another requirement in running a regression model is the assumption of normality and sphericity of the error term. If the assumption of normality is violated or outliers are present, then the linear regression goodness-of-fit test may not be the most powerful or informative test available, and this could mean the difference between detecting a linear fit or not. If the errors are not independent and not normally distributed, it may indicate that the data might be autocorrelated or suffer from nonlinearities or other more destructive errors. Independence of the errors can also be detected in the heteroskedasticity tests (Figure 5.25).

The Normality test on the errors performed is a nonparametric test, which makes no assumptions about the specific shape of the population, from which the samples are drawn, allowing for smaller sample data sets to be analyzed. This test evaluates the null hypothesis of whether the sample errors were drawn from a normally distributed population, versus an alternate hypothesis that the data sample is not normally distributed. If the calculated D-statistic is greater than or equal to the D-critical values at various significance values, then reject the null hypothesis and accept the alternate hypothesis (the errors are not normally distributed). Otherwise, if the D-statistic is less than the D-critical value, do not reject the null hypothesis (the errors are normally distributed). The Normality test relies on two cumulative frequencies: one derived from the sample data set and the second from a theoretical distribution based on the mean and standard deviation of the sample data.

Test Result

Regression Error Average	0.00
Standard Deviation of Errors	141.83
D Statistic	0.1036
D Critical at 1%	0.1138
D Critical at 5%	0.1225
D Critical at 10%	0.1458

Null Hypothesis: The errors are normally distributed.

Conclusion: The errors are normally distributed at the 1% alpha level.

Errors	Relative Frequency	Observed	Expected	O-E
-219.04	0.02	0.02	0.0612	-0.0412
-202.53	0.02	0.04	0.0766	-0.0366
-186.04	0.02	0.06	0.0948	-0.0348
-174.17	0.02	0.08	0.1097	-0.0297
-162.13	0.02	0.10	0.1265	-0.0265
-161.62	0.02	0.12	0.1272	-0.0072
-160.39	0.02	0.14	0.1291	0.0109
-145.40	0.02	0.16	0.1526	0.0074
-138.92	0.02	0.18	0.1637	0.0163
-133.81	0.02	0.20	0.1727	0.0273
-120.76	0.02	0.22	0.1973	0.0227
-120.12	0.02	0.24	0.1985	0.0415

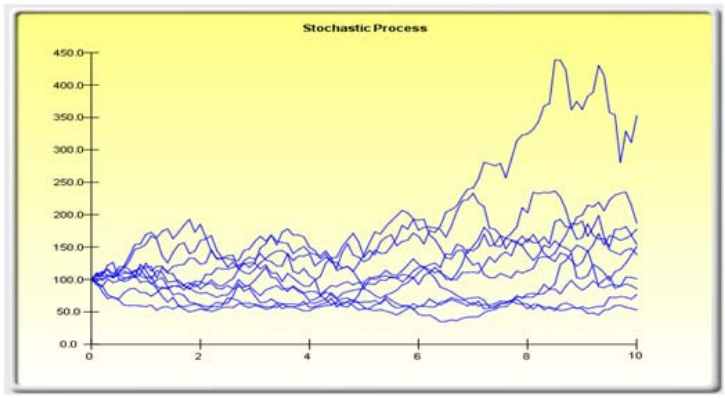
Figure 5.25 – Test for Normality of Errors

Sometimes, certain types of time-series data cannot be modeled using any other methods except for a stochastic process, because the underlying events are stochastic in nature. For instance, you cannot adequately model and forecast stock prices, interest rates, price of oil, and other commodity prices using a simple regression model because these variables are highly uncertain and volatile, and they do not follow a predefined static rule of behavior; in other words, the process is not stationary. Stationarity is checked using the Runs Test function, while another visual clue is found in the autocorrelation report (the ACF tends to decay slowly). A stochastic process is a sequence of events or paths generated by probabilistic laws. That is, random events can occur over time but are governed by specific statistical and probabilistic rules. The main stochastic processes include random walk or Brownian motion, mean-reversion, and jump-diffusion. These processes can be used to forecast a multitude of variables that seemingly follow random trends but restricted by probabilistic laws. The process-generating equation is known in advance but the actual results generated are unknown (Figure 5.26).

The Random Walk Brownian Motion process can be used to forecast stock prices, prices of commodities, and other stochastic time-series data given a drift or growth rate and volatility around the drift path. The Mean-Reversion process can be used to reduce the fluctuations of the Random Walk process by allowing the path to target a long-term value, making it useful for forecasting time-series variables that have a long-term rate such as interest rates and inflation rates (these are long-term target rates by regulatory authorities or the market). The Jump-Diffusion process is useful for forecasting time-series data when the variable can occasionally exhibit random jumps, such as oil prices or price of electricity (discrete exogenous event shocks can make prices jump up or down). These processes can also be mixed and matched as required.

A note of caution is required here. The stochastic parameters calibration shows all the parameters for all processes and does not distinguish which process is better and which is worse or which process is more appropriate to use. It is up to the user to make this determination. For instance, if we see a 283% reversion rate, chances are, a mean-reversion process is inappropriate; or a very high jump rate of, say, 100% most probably means that a jump-diffusion process is probably not appropriate; and so forth. Further, the analysis cannot determine what the variable is and what the data source is. For instance, is the raw data from historical stock prices or is it the historical prices of electricity or inflation rates or the molecular motion of subatomic particles, and so forth. Only the user would know about the raw data, and, hence, using a priori knowledge and theory, be able to pick the correct process to

use (e.g., stock prices tend to follow a Brownian motion random walk, whereas inflation rates follow a mean-reversion process; or a jump-diffusion process is more appropriate should you be forecasting the price of electricity).



Statistical Summary

The following are the estimated parameters for a stochastic process given the data provided. It is up to you to determine if the probability of fit (similar to a goodness-of-fit computation) is sufficient to warrant the use of a stochastic process forecast, and if so, whether it is a random walk, mean-reversion, or a jump-diffusion model, or combinations thereof. In choosing the right stochastic process model, you will have to rely on past experiences and *a priori* economic and financial expectations of what the underlying data set is best represented by. These parameters can be entered into a stochastic process forecast (**Simulation** | **Forecasting** | **Stochastic Processes**).

Periodic

Drift Rate	*-1.48%*	*Reversion Rate*	*283.89%*	*Jump Rate*	*20.41%*
Volatility	*88.84%*	*Long-Term Value*	*327.72*	*Jump Size*	*237.89*

Probability of stochastic model fit: 46.48%
A high fit means a stochastic model is better than conventional models.

Runs	*20*	*Standard Normal*	*-1.7321*
Positive	*25*	*P-Value (1-tail)*	*0.0416*
Negative	*25*	*P-Value (2-tail)*	*0.0833*
Expected Run	*26*		

A low p-value (below 0.10, 0.05, 0.01) means that the sequence is not random and hence suffers from stationarity problems, and an ARIMA model might be more appropriate. Conversely, higher p-values indicate randomness and stochastic process models might be appropriate.

Figure 5.26 – Stochastic Process Parameter Estimation

Multicollinearity exists when there is a linear relationship between the independent variables. When this occurs, the regression equation cannot be estimated at all. In near collinearity situations, the estimated regression equation will be biased and provide inaccurate results. This situation is especially true when a stepwise regression approach is used, where the statistically significant independent variables will be thrown out of the regression mix earlier than expected, resulting in a regression equation that is neither efficient nor accurate. One quick test of the presence of multicollinearity in a multiple regression equation is that the R-squared value is relatively high, while the t-statistics are relatively low.

Another quick test is to create a correlation matrix between the independent. A high cross-correlation indicates a potential for autocorrelation. The rule of thumb is that a correlation with an absolute value greater than 0.75 is indicative of severe multicollinearity.

Correlation Matrix

CORRELATION	X2	X3	X4	X5
X1	0.333	0.959	0.242	0.237
X2	1.000	0.349	0.319	0.120
X3		1.000	0.196	0.227
X4			1.000	0.290

Variance Inflation Factor

VIF	X2	X3	X4	X5
X1	1.12	**12.46**	1.06	1.06
X2	N/A	1.14	1.11	1.01
X3		N/A	1.04	1.05
X4			N/A	1.09

Figure 5.27 – Multicollinearity Errors

The Correlation Matrix lists the Pearson's Product Moment Correlations (commonly referred to as the Pearson's R) between variable pairs. The correlation coefficient ranges between –1.0 and + 1.0 inclusive. The sign indicates the direction of association between the variables, while the coefficient indicates the magnitude or strength of association. The Pearson's R only measures a linear relationship and is less effective in measuring nonlinear relationships.

To test whether the correlations are significant, a two-tailed hypothesis test is performed and the resulting p-value(s) is listed. In Figure 5.27 (top), P-values less than 0.10, 0.05, and 0.01 are highlighted in blue to indicate statistical significance. In other words, a p-value for a correlation pair that is less than a given significance value is statistically significantly different from zero, indicating that there is significant a linear relationship between the two variables.

The Pearson's R between two variables (x and y) is related to the covariance (cov) measure, where: $R_{x,y} = \frac{COV_{x,y}}{s_x s_y}$. The benefit of dividing the covariance by the product of the two variables' standard deviation (s) is that the resulting correlation coefficient is bounded between –1.0 and +1.0 inclusive. This makes the correlation a good relative measure to compare among different variables (particularly with different units and magnitude). The Spearman rank-based nonparametric correlation is also included in the report. The Spearman's R is related to the Pearson's R in that the data is first ranked and then correlated. The rank correlation provides a better estimate of the relationship between two variables when one or both of them is nonlinear.

It must be stressed that a significant correlation does not imply causation. Associations between variables in no way imply that the change of one variable causes another variable to change. When two variables that are moving independently of each other but in a related path, they may be correlated but their relationship might be spurious (e.g., a correlation between sunspots and the stock market might be strong, but one can surmise that there is no causality and that this relationship is purely spurious). Another test for multicollinearity is the use of the variance inflation factor (VIF), obtained by regressing each independent variable to all the other independent variables, obtaining the R-squared value, and calculating the VIF. A VIF exceeding 2.0 can be considered as severe multicollinearity. A VIF exceeding 10.0 indicates destructive multicollinearity (Figure 5.27, bottom).

5.9 Statistical Analysis Tool

Another very powerful tool in Risk Simulator is the Statistical Analysis tool, which determines the statistical properties of the data. The diagnostics run include checking the data for various statistical properties, from basic descriptive statistics to testing for and calibrating the stochastic properties of the data.

Procedure

- Open the example model (*Risk Simulator | Examples | Statistical Analysis*), go to the Data worksheet, and select the data including the variable names (cells *C5:E55*).
- Click on *Risk Simulator | Tools | Statistical Analysis* (Figure 5.28).
- Check the data type, whether the data selected are from a single variable or multiple variables arranged in rows. In our example, we assume that the data areas selected are from multiple variables. Click *OK* when finished.
- Choose the statistical tests you wish to perform. The suggestion (and by default) is to choose all the tests. Click *OK* when finished (Figure 5.29).

Spend some time going through the reports generated to get a better understanding of the statistical tests performed (sample reports are shown in Figures 5.30 through 5.33).

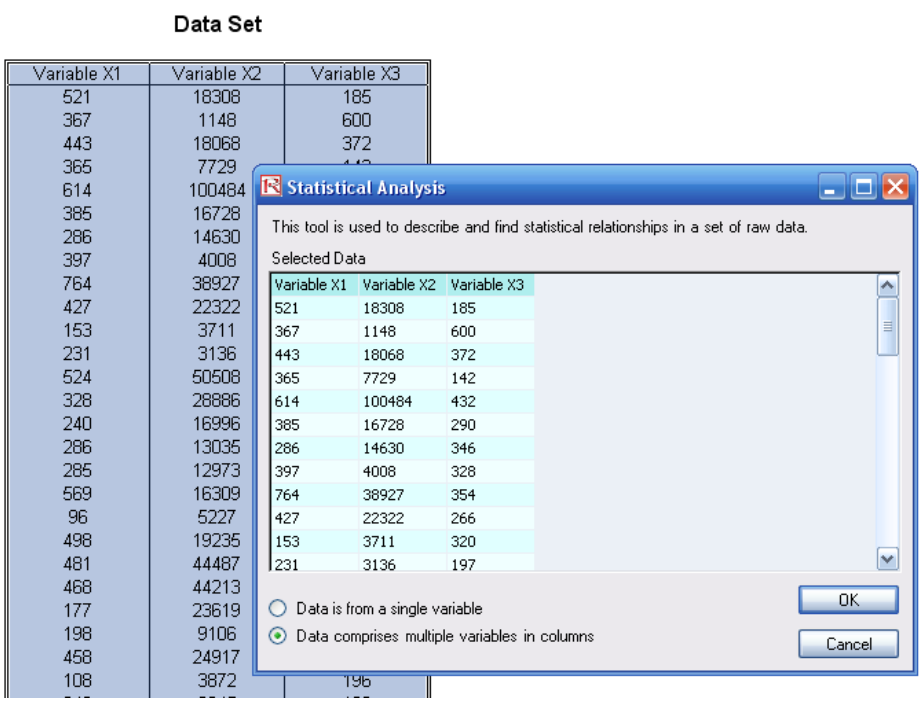


Figure 5.28 – Running the Statistical Analysis Tool

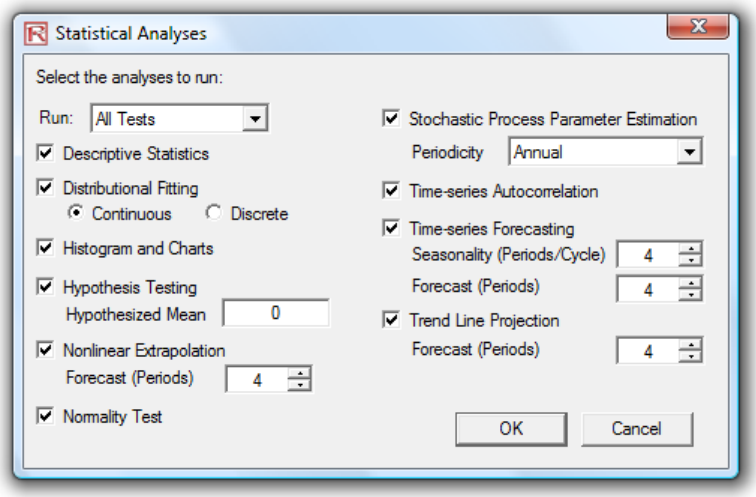


Figure 5.29 – Statistical Tests

Descriptive Statistics

Analysis of Statistics

Almost all distributions can be described within 4 moments (some distributions require one moment, while others require two moments, and so forth). Descriptive statistics quantitatively capture these moments. The first moment describes the location of a distribution (i.e., mean, median, and mode) and is interpreted as the expected value, expected returns, or the average value of occurrences.

The Arithmetic Mean calculates the average of all occurrences by summing up all of the data points and dividing them by the number of points. The Geometric Mean is calculated by taking the power root of the products of all the data points and requires them to all be positive. The Geometric Mean is more accurate for percentages or rates that fluctuate significantly. For example, you can use Geometric Mean to calculate average growth rate given compound interest with variable rates. The Trimmed Mean calculates the arithmetic average of the data set after the extreme outliers have been trimmed. As averages are prone to significant bias when outliers exist, the Trimmed Mean reduces such bias in skewed distributions.

The Standard Error of the Mean calculates the error surrounding the sample mean. The larger the sample size, the smaller the error such that for an infinitely large sample size, the error approaches zero, indicating that the population parameter has been estimated. Due to sampling errors, the 95% Confidence Interval for the Mean is provided. Based on an analysis of the sample data points, the actual population mean should fall between these Lower and Upper Intervals for the Mean.

Median is the data point where 50% of all data points fall above this value and 50% below this value. Among the three first moment statistics, the median is least susceptible to outliers. A symmetrical distribution has the Median equal to the Arithmetic Mean. A skewed distribution exists when the Median is far away from the Mean. The Mode measures the most frequently occurring data point.

Minimum is the smallest value in the data set while Maximum is the largest value. Range is the difference between the Maximum and Minimum values.

The second moment measures a distribution's spread or width, and is frequently described using measures such as Standard Deviations, Variances, Quartiles, and Inter-Quartile Ranges. Standard Deviation indicates the average deviation of all data points from their mean. It is a popular measure as is associated with risk (higher standard deviations mean a wider distribution, higher risk, or wider dispersion of data points around the mean) and its units are identical to original data set's. The Sample Standard Deviation differs from the Population Standard Deviation in that the former uses a degree of freedom correction to account for small sample sizes. Also, Lower and Upper Confidence Intervals are provided for the Standard Deviation and the true population standard deviation falls within this interval. If your data set covers every element of the population, use the Population Standard Deviation instead. The two Variance measures are simply the squared values of the standard deviations.

The Coefficient of Variability is the standard deviation of the sample divided by the sample mean, proving a unit-free measure of dispersion that can be compared across different distributions (you can now compare distributions of values denominated in millions of dollars with one in billions of dollars, or meters and kilograms, etc.). The First Quartile measures the 25th percentile of the data points when arranged from its smallest to largest value. The Third Quartile is the value of the 75th percentile data point. Sometimes quartiles are used as the upper and lower ranges of a distribution as it truncates the data set to ignore outliers. The Inter-Quartile Range is the difference between the third and first quartiles, and is often used to measure the width of the center of a distribution.

Skewness is the third moment in a distribution. Skewness characterizes the degree of asymmetry of a distribution around its mean. Positive skewness indicates a distribution with an asymmetric tail extending toward more positive values. Negative skewness indicates a distribution with an asymmetric tail extending toward more negative values.

Kurtosis characterizes the relative peakedness or flatness of a distribution compared to the normal distribution. It is the fourth moment in a distribution. A positive Kurtosis value indicates a relatively peaked distribution. A negative kurtosis indicates a relatively flat distribution. The Kurtosis measured here has been centered to zero (certain other kurtosis measures are centered around 3.0). While both are equally valid, centering across zero makes the interpretation simpler. A high positive Kurtosis indicates a peaked distribution around its center and leptokurtic or fat tails. This indicates a higher probability of extreme events (e.g., catastrophic events, terrorist attacks, stock market crashes) than is predicted in a normal distribution.

Summary Statistics

Statistics	*Variable X1*		
Observations	*50.0000*	*Standard Deviation (Sample)*	*172.9140*
Arithmetic Mean	*331.9200*	*Standard Deviation (Population)*	*171.1761*
Geometric Mean	*281.3247*	*Lower Confidence Interval for Standard Deviation*	*148.6090*
Trimmed Mean	*325.1739*	*Upper Confidence Interval for Standard Deviation*	*207.7947*
Standard Error of Arithmetic Mean	*24.4537*	*Variance (Sample)*	*29899.2588*
Lower Confidence Interval for Mean	*283.0125*	*Variance (Population)*	*29301.2736*
Upper Confidence Interval for Mean	*380.8275*	*Coefficient of Variability*	*0.5210*
Median	*307.0000*	*First Quartile (Q1)*	*204.0000*
Mode	*47.0000*	*Third Quartile (Q3)*	*441.0000*
Minimum	*764.0000*	*Inter-Quartile Range*	*237.0000*
Maximum	*717.0000*	*Skewness*	*0.4838*
Range		*Kurtosis*	*-0.0952*

Figure 5.30 – Sample Statistical Analysis Tool Report

Hypothesis Test (t-Test on the Population Mean of One Variable)

Statistical Summary

Statistics from Dataset:		*Calculated Statistics:*	
Observations	*50*	*t-Statistic*	*13.5734*
Sample Mean	*331.92*	*P-Value (right-tail)*	***0.0000***
Sample Standard Deviation	*172.91*	*P-Value (left-tailed)*	***1.0000***
		P-Value (two-tailed)	***0.0000***
User Provided Statistics:			
		Null Hypothesis (Ho):	*μ = Hypothesized Mean*
Hypothesized Mean	*0.00*	*Alternate Hypothesis (Ha):*	*μ < > Hypothesized Mean*

Notes: "<>" denotes "greater than" for right-tail, "less than" for left-tail, or "not equal to" for two-tail hypothesis tests.

Hypothesis Testing Summary

The one-variable t-test is appropriate when the population standard deviation is not known but the sampling distribution is assumed to be approximately normal (the t-test is used when the sample size is less than 30 but is also appropriate and in fact, provides more conservative results with larger data sets). This t-test can be applied to three types of hypothesis tests: a two-tailed test, a right-tailed test, and a left-tailed test. All three tests and their respective results are listed below for your reference.

Two-Tailed Hypothesis Test

A two-tailed hypothesis tests the null hypothesis Ho such that the population mean is statistically identical to the hypothesized mean. The alternative hypothesis is that the real population mean is statistically different from the hypothesized mean when tested using the sample dataset. Using a t-test, if the computed p-value is less than a specified significance amount (typically 0.10, 0.05, or 0.01), this means that the population mean is statistically significantly different than the hypothesized mean at 10%, 5% and 1% significance value (or at the 90%, 95%, and 99% statistical confidence). Conversely, if the p-value is higher than 0.10, 0.05, or 0.01, the population mean is statistically identical to the hypothesized mean and any differences are due to random chance.

Right-Tailed Hypothesis Test

A right-tailed hypothesis tests the null hypothesis Ho such that the population mean is statistically less than or equal to the hypothesized mean. The alternative hypothesis is that the real population mean is statistically greater than the hypothesized mean when tested using the sample dataset. Using a t-test, if the p-value is less than a specified significance amount (typically 0.10, 0.05, or 0.01), this means that the population mean is statistically significantly greater than the hypothesized mean at 10%, 5% and 1% significance value (or 90%, 95%, and 99% statistical confidence). Conversely, if the p-value is higher than 0.10, 0.05, or 0.01, the population mean is statistically similar or less than the hypothesized mean.

Left-Tailed Hypothesis Test

A left-tailed hypothesis tests the null hypothesis Ho such that the population mean is statistically greater than or equal to the hypothesized mean. The alternative hypothesis is that the real population mean is statistically less than the hypothesized mean when tested using the sample dataset. Using a t-test, if the p-value is less than a specified significance amount (typically 0.10, 0.05, or 0.01), this means that the population mean is statistically significantly less than the hypothesized mean at 10%, 5%, and 1% significance value (or 90%, 95%, and 99% statistical confidence). Conversely, if the p-value is higher than 0.10, 0.05, or 0.01, the population mean is statistically similar or greater than the hypothesized mean and any differences are due ti random chance.

Because the t-test is more conservative and does not require a known population standard deviation as in the Z-test, we only use this t-test.

Figure 5.31 – Sample Statistical Analysis Tool Report (Hypothesis Testing of One Variable)

Test for Normality

The Normality test is a form of nonparametric test, which makes no assumptions about the specific shape of the population from which the sample is drawn, allowing for smaller sample data sets to be analyzed. This test evaluates the null hypothesis of whether the data sample was drawn from a normally distributed population, versus an alternate hypothesis that the data sample is not normally distributed. If the calculated p-value is less than or equal to the alpha significance value then reject the null hypothesis and accept the alternate hypothesis. Otherwise, if the p-value is higher than the alpha significance value, do not reject the null hypothesis. This test relies on two cumulative frequencies: one derived from the sample data set, the second from a theoretical distribution based on the mean and standard deviation of the sample data. An alternative to this test is the Chi-Square test for normality. The Chi-Square test requires more data points to run compared to the Normality test used here.

Test Result

Data Average	*331.92*
Standard Deviation	*172.91*
D Statistic	*0.0859*
D Critical at 1%	*0.1150*
D Critical at 5%	*0.1237*
D Critical at 10%	*0.1473*

Null Hypothesis: The data is normally distributed.

Conclusion: The sample data is normally distributed at the 1% alpha level.

Data	*Relative Frequency*	*Observed*	*Expected*	*O-E*
47.00	*0.02*	*0.02*	*0.0497*	*-0.0297*
68.00	*0.02*	*0.04*	*0.0635*	*-0.0235*
87.00	*0.02*	*0.06*	*0.0783*	*-0.0183*
96.00	*0.02*	*0.08*	*0.0862*	*-0.0062*
102.00	*0.02*	*0.10*	*0.0918*	*0.0082*
108.00	*0.02*	*0.12*	*0.0977*	*0.0223*
114.00	*0.02*	*0.14*	*0.1038*	*0.0362*
127.00	*0.02*	*0.16*	*0.1180*	*0.0420*
153.00	*0.02*	*0.18*	*0.1504*	*0.0296*
177.00	*0.02*	*0.20*	*0.1851*	*0.0149*
186.00	*0.02*	*0.22*	*0.1994*	*0.0206*
188.00	*0.02*	*0.24*	*0.2026*	*0.0374*
198.00	*0.02*	*0.26*	*0.2193*	*0.0407*
222.00	*0.02*	*0.28*	*0.2625*	*0.0175*
231.00	*0.02*	*0.30*	*0.2797*	*0.0203*
240.00	*0.02*	*0.32*	*0.2975*	*0.0225*
246.00	*0.02*	*0.34*	*0.3096*	*0.0304*
251.00	*0.02*	*0.36*	*0.3199*	*0.0401*
265.00	*0.02*	*0.38*	*0.3494*	*0.0306*
280.00	*0.02*	*0.40*	*0.3820*	*0.0180*
285.00	*0.02*	*0.42*	*0.3931*	*0.0269*
286.00	*0.04*	*0.46*	*0.3953*	*0.0647*
291.00	*0.02*	*0.48*	*0.4065*	*0.0735*
303.00	*0.02*	*0.50*	*0.4336*	*0.0664*
311.00	*0.02*	*0.52*	*0.4519*	*0.0681*

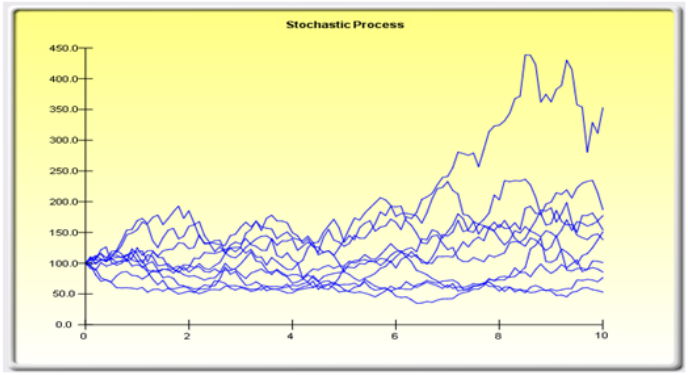
Figure 5.32 – Sample Statistical Analysis Tool Report (Normality Test)

Stochastic Process - Parameter Estimations

Statistical Summary

A stochastic process is a sequence of events or paths generated by probabilistic laws. That is, random events can occur over time but are governed by specific statistical and probabilistic rules. The main stochastic processes include Random Walk or Brownian Motion, Mean-Reversion, and Jump-Diffusion. These processes can be used to forecast a multitude of variables that seemingly follow random trends but yet are restricted by probabilistic laws. The process-generating equation is known in advance but the actual results generated is unknown.

The Random Walk Brownian Motion process can be used to forecast stock prices, prices of commodities, and other stochastic time-series data given a drift or growth rate and a volatility around the drift path. The Mean-Reversion process can be used to reduce the fluctuations of the Random Walk process by allowing the path to target a long-term value, making it useful for forecasting time-series variables that have a long-term rate such as interest rates and inflation rates (these are long-term target rates by regulatory authorities or the market). The Jump-Diffusion process is useful for forecasting time-series data when the variable can occasionally exhibit random jumps, such as oil prices or price of electricity (discrete exogenous event shocks can make prices jump up or down). Finally, these three stochastic processes can be mixed and matched as required.



Statistical Summary

The following are the estimated parameters for a stochastic process given the data provided. It is up to you to determine if the probability of fit (similar to a goodness-of-fit computation) is sufficient to warrant the use of a stochastic process forecast, and if so, whether it is a random walk, mean-reversion, or a jump-diffusion model, or combinations thereof. In choosing the right stochastic process model, you will have to rely on past experiences and a priori economic and financial expectations of what the underlying data set is best represented by. These parameters can be entered into a stochastic process forecast (**Simulation | Forecasting | Stochastic Processes**).

(Annualized)

Drift Rate	*5.86%*	*Reversion Rate*	*N/A*	*Jump Rate*	*16.33%*
Volatility	*7.04%*	*Long-Term Value*	*N/A*	*Jump Size*	*21.33*

Probability of stochastic model fit: *4.63%*

Figure 5.33 – Sample Statistical Analysis Tool Report (Stochastic Parameter Estimation)

5.10 Distributional Analysis Tool

The Distributional Analysis tool is a statistical probability tool in Risk Simulator that is useful in a variety of settings. It can be used to compute the probability density function (PDF), which is also called the probability mass function (PMF) for discrete distributions (these terms are used interchangeably), where given some distribution and its parameters, we can determine the probability of occurrence given some outcome x. In addition, the cumulative distribution function (CDF) can be computed, which is the sum of the PDF values up to this x value. Finally, the inverse cumulative distribution function (ICDF) is used to compute the value x given the cumulative probability of occurrence.

This tool is accessible via *Risk Simulator | Tools | Distributional Analysis*. As an example of its use, Figure 5.34 shows the computation of a binomial distribution (i.e., a distribution with two outcomes, such as the tossing of a coin, where the outcome is either Head or Tail, with some prescribed probability of heads and tails). Suppose we toss a coin two times. Setting the outcome Head as a success, we use the binomial distribution with Trials = 2 (tossing the coin twice) and Probability = 0.50 (the probability of success, of getting Heads). Selecting the PDF

and setting the range of values x as from 0 to 2 with a step size of 1 (this means we are requesting the values 0, 1, 2 for x), the resulting probabilities, as well as the theoretical four moments of the distribution, are provided in tabular and in graphical formats. As the outcomes of the coin toss are Heads-Heads, Tails-Tails, Heads-Tails, and Tails-Heads, the probability of getting exactly no Heads is 25%, one Head is 50%, and two Heads is 25%. Similarly, we can obtain the exact probabilities of tossing the coin, say, 20 times, as seen in Figure 5.35.

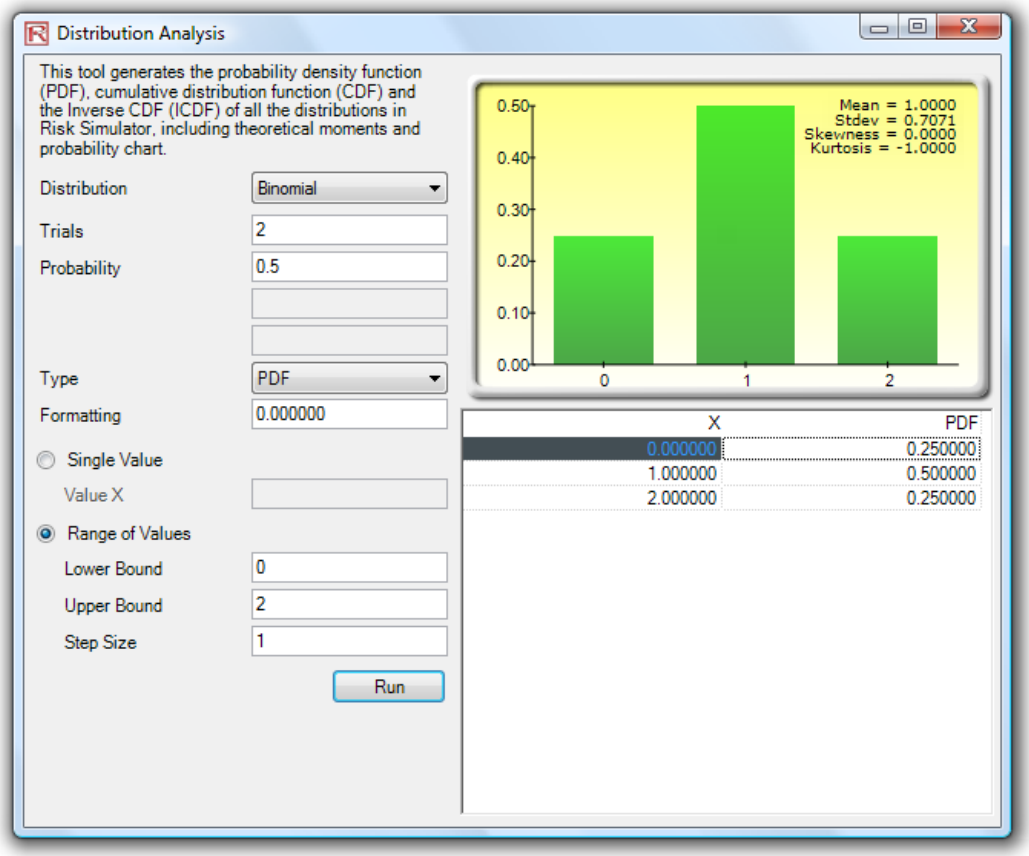


Figure 5.34 – Distributional Analysis Tool (Binomial Distribution with 2 Trials)

Figure 5.36 shows the same binomial distribution for 20 trials, but now the CDF is computed. The CDF is simply the sum of the PDF values up to the point x. For instance, in Figure 5.35, we see that the probabilities of 0, 1, and 2 are 0.000001, 0.000019, and 0.000181, whose sum is 0.000201, which is the value of the CDF at x = 2 in Figure 5.36. Whereas the PDF computes the probabilities of getting exactly 2 heads, the CDF computes the probability of getting no more than 2 heads or up to 2 heads (or probabilities of 0, 1, and 2 heads). Taking the complement (i.e., 1 – 0.00021) obtains 0.999799, or 99.9799%, which is the probability of getting at least 3 heads or more.

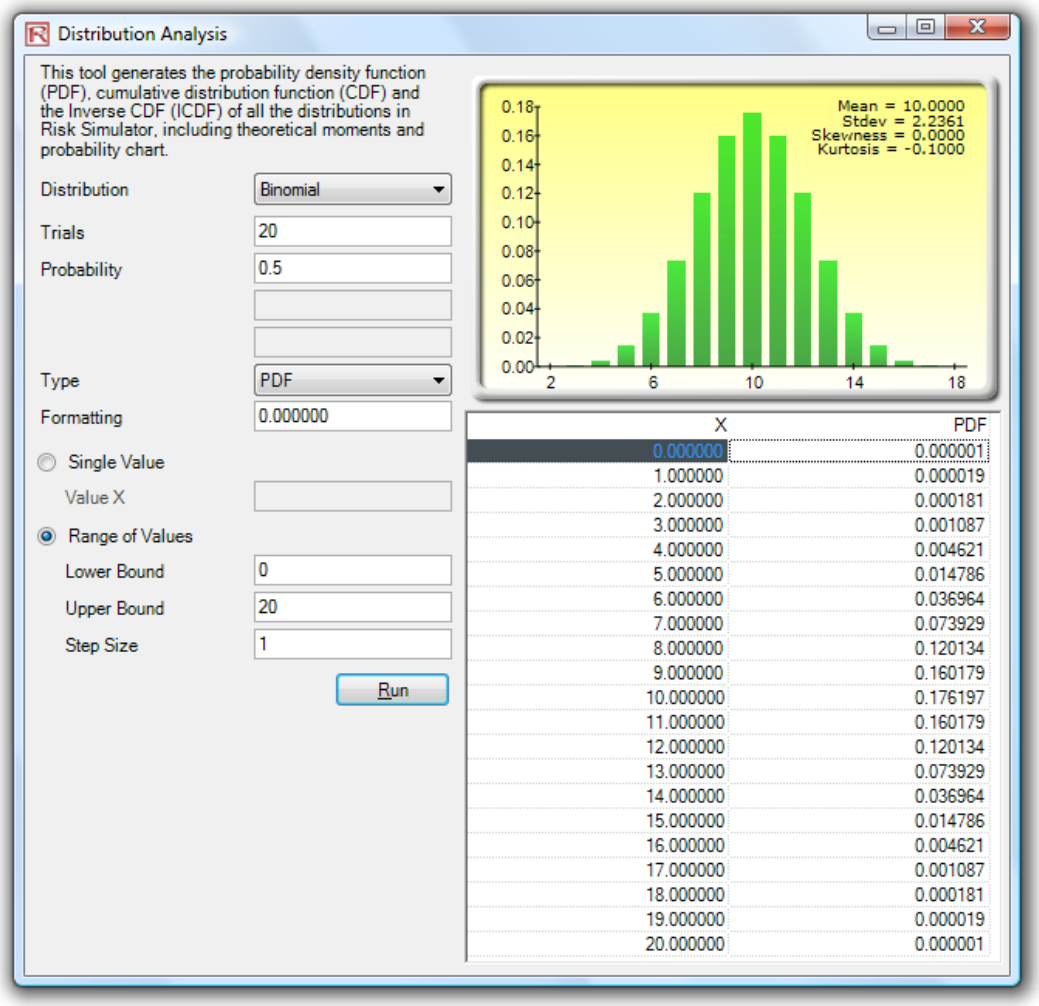


Figure 5.35 – Distributional Analysis Tool (Binomial Distribution with 20 Trials)

Using this Distributional Analysis tool in Risk Simulator, even more advanced distributions can be analyzed, such as the gamma, beta, negative binomial, and many others. As further example of the tool's use in a continuous distribution and the ICDF functionality, Figure 5.37 shows the standard normal distribution (normal distribution with a mean of zero and standard deviation of one), where we apply the ICDF to find the value of x that corresponds to the cumulative probability of 97.50% (CDF). That is, a one-tail CDF of 97.50% is equivalent to a two-tail 95% confidence interval (there is a 2.50% probability in the right tail and 2.50% in the left tail, leaving 95% in the center or confidence interval area, which is equivalent to a 97.50% area for one tail). The result is the familiar Z-score of 1.96. Therefore, using this Distributional Analysis tool, the standardized scores for other distributions and the exact and cumulative probabilities of other distributions can all be obtained quickly and easily.

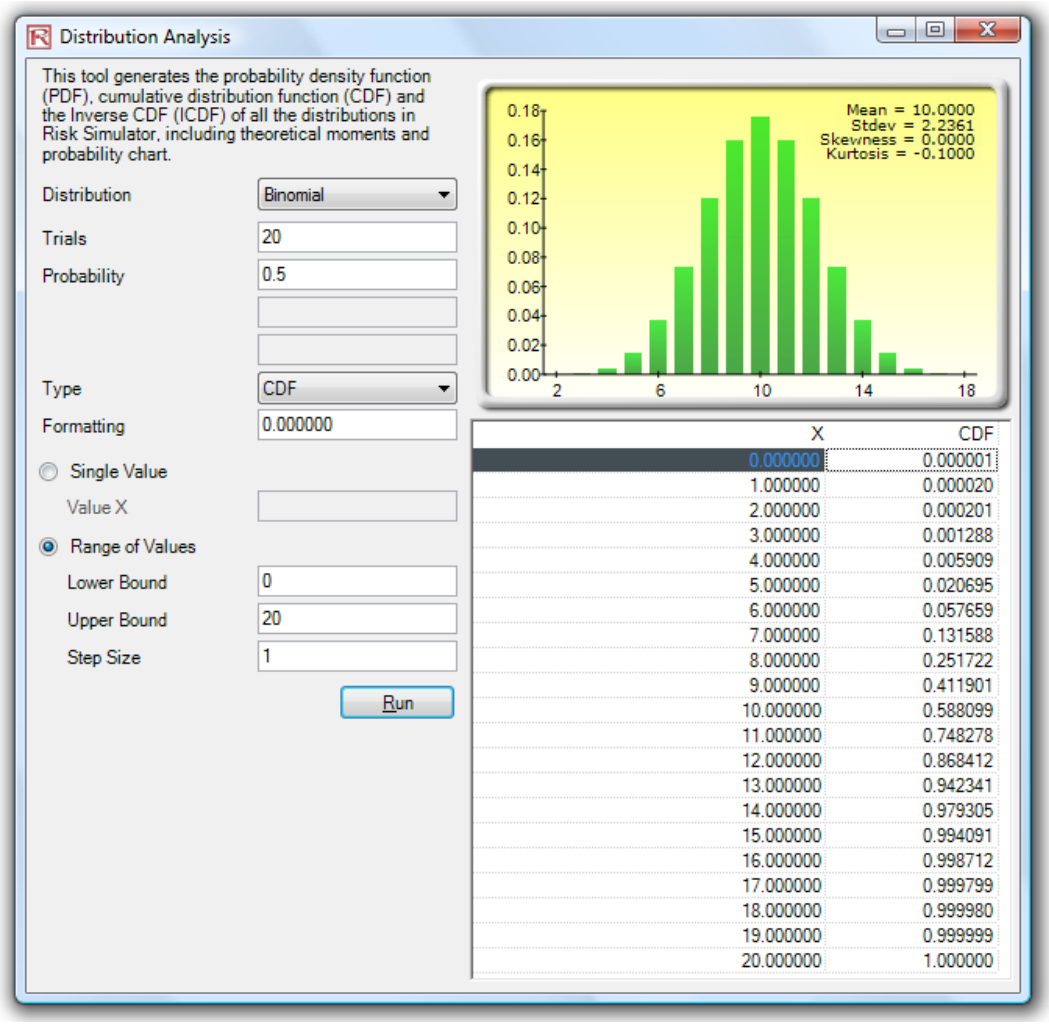


Figure 5.36 – Distributional Analysis Tool (Binomial Distribution's CDF with 20 Trials)

5.11 Scenario Analysis Tool

The Scenario Analysis tool in Risk Simulator allows you to run multiple scenarios quickly and effortlessly by changing one or two input parameters to determine the output of a variable. Figure 5.38 illustrates how this tool works on the discounted cash flow sample model (Model 7 in Risk Simulator's Example Models folder). In this example, cell G6 (net present value) is selected as the output of interest, whereas cells C9 (effective tax rate) and C12 (product price) are selected as inputs to perturb. You can set the starting and ending values to test, as well as the step size, or the number of steps, to run between these starting and ending values. The result is a scenario analysis table (Figure 5.39), where the row and column headers are the two input variables and the body of the table shows the net present values.

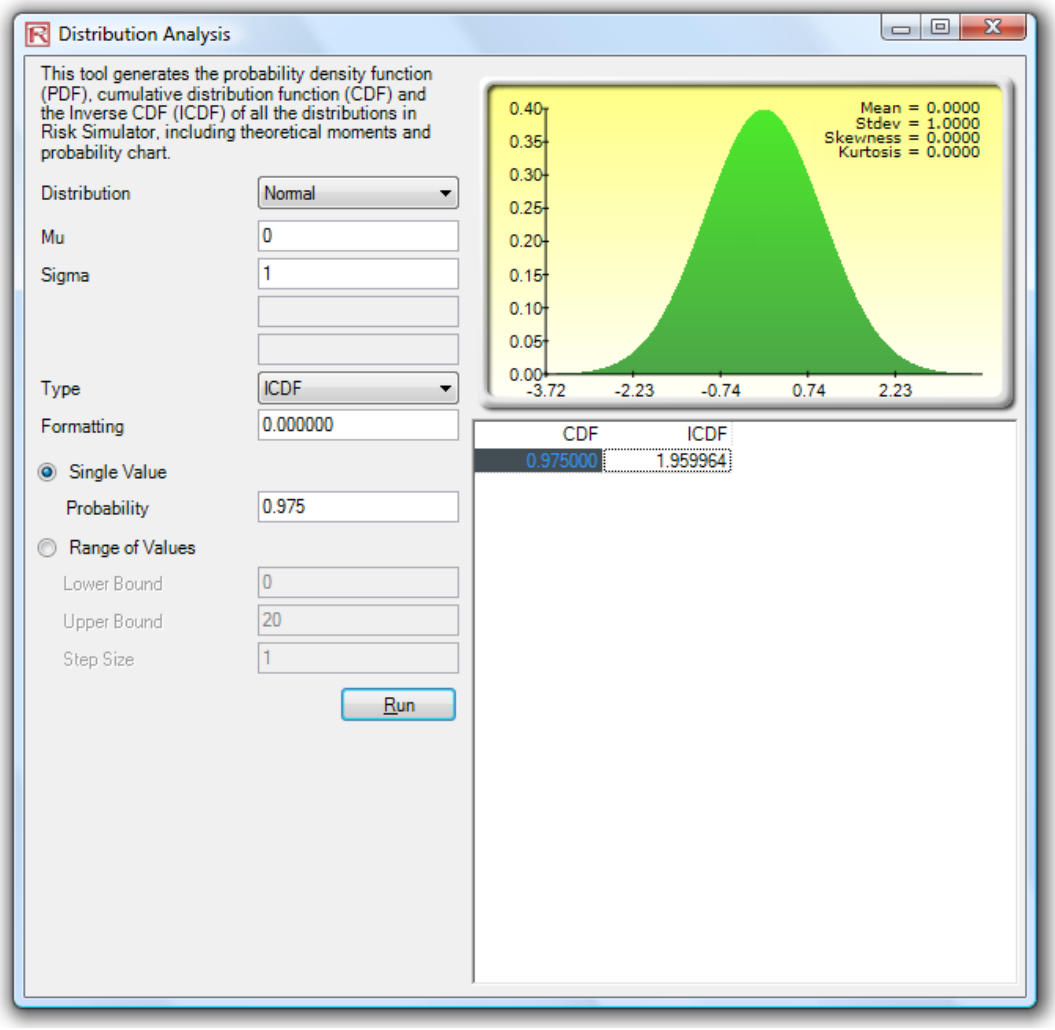


Figure 5.37 – Distributional Analysis Tool (Normal Distribution's ICDF and Z-Score)

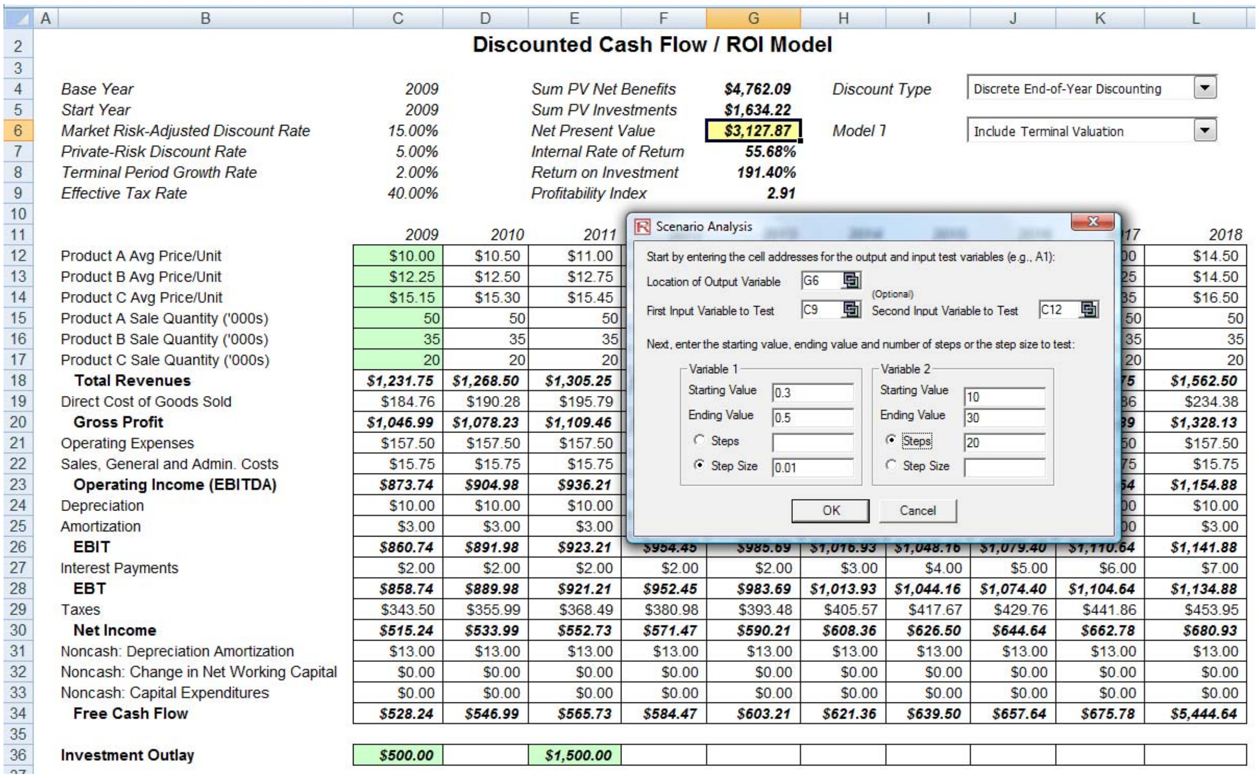


Figure 5.38 – Scenario Analysis Tool

SCENARIO ANALYSIS TABLE

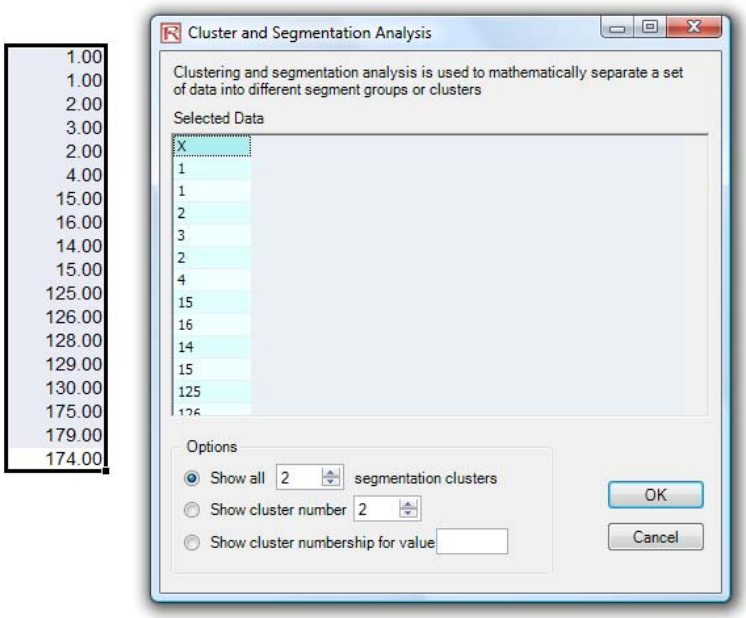
Output Variable:	G6	**Initial Base Case Value:**		$3,127.87					
Column Variable:	C12	**Min:**	10 **Max:**	30 **Steps:**	20	**Stepsize:** ---		**Initial Base Case Value:**	$10.00
Row Variable:	C9	**Min:**	0.3 **Max:**	0.5 **Steps:**	---	**Stepsize:**	0.01	**Initial Base Case Value:**	40.00%

	$10.00	**$11.00**	**$12.00**	**$13.00**	**$14.00**	**$15.00**	**$16.00**	**$17.00**	**$18.00**	**$19.00**	**$20.00**	**$21.00**	**$22.00**	**$23.00**	**$24.00**	**$25.00**	**$26.00**	**$27.00**	**$28.00**	**$29.00**	**$30.00**
30.00%	$3,904.83	$4,134.43	$4,364.04	$4,593.64	$4,823.24	$5,052.84	$5,282.44	$5,512.04	$5,741.64	$5,971.24	$6,200.85	$6,430.45	$6,660.05	$6,889.65	$7,119.25	$7,348.85	$7,578.45	$7,808.05	$8,037.65	$8,267.26	$8,496.86
31.00%	$3,827.14	$4,053.46	$4,279.78	$4,506.10	$4,732.42	$4,958.74	$5,185.06	$5,411.39	$5,637.71	$5,864.03	$6,090.35	$6,316.67	$6,542.99	$6,769.31	$6,995.63	$7,221.96	$7,448.28	$7,674.60	$7,900.92	$8,127.24	$8,353.56
32.00%	$3,749.44	$3,972.48	$4,195.52	$4,418.56	$4,641.61	$4,864.65	$5,087.69	$5,310.73	$5,533.77	$5,756.81	$5,979.85	$6,202.89	$6,425.94	$6,648.98	$6,872.02	$7,095.06	$7,318.10	$7,541.14	$7,764.18	$7,987.22	$8,210.26
33.00%	$3,671.75	$3,891.51	$4,111.27	$4,331.03	$4,550.79	$4,770.55	$4,990.31	$5,210.07	$5,429.83	$5,649.60	$5,869.36	$6,089.12	$6,308.88	$6,528.64	$6,748.40	$6,968.16	$7,187.92	$7,407.68	$7,627.45	$7,847.21	$8,066.97
34.00%	$3,594.05	$3,810.53	$4,027.01	$4,243.49	$4,459.97	$4,676.45	$4,892.94	$5,109.42	$5,325.90	$5,542.38	$5,758.86	$5,975.34	$6,191.82	$6,408.30	$6,624.79	$6,841.27	$7,057.75	$7,274.23	$7,490.71	$7,707.19	$7,923.67
35.00%	$3,516.35	$3,729.55	$3,942.76	$4,155.96	$4,369.16	$4,582.36	$4,795.56	$5,008.76	$5,221.96	$5,435.16	$5,648.36	$5,861.57	$6,074.77	$6,287.97	$6,501.17	$6,714.37	$6,927.57	$7,140.77	$7,353.97	$7,567.17	$7,780.38
36.00%	$3,438.66	$3,648.58	$3,858.50	$4,068.42	$4,278.34	$4,488.26	$4,698.18	$4,908.10	$5,118.03	$5,327.95	$5,537.87	$5,747.79	$5,957.71	$6,167.63	$6,377.55	$6,587.47	$6,797.39	$7,007.32	$7,217.24	$7,427.16	$7,637.08
37.00%	$3,360.96	$3,567.60	$3,774.24	$3,980.88	$4,187.53	$4,394.17	$4,600.81	$4,807.45	$5,014.09	$5,220.73	$5,427.37	$5,634.01	$5,840.65	$6,047.30	$6,253.94	$6,460.58	$6,667.22	$6,873.86	$7,080.50	$7,287.14	$7,493.78
38.00%	$3,283.27	$3,486.63	$3,689.99	$3,893.35	$4,096.71	$4,300.07	$4,503.43	$4,706.79	$4,910.15	$5,113.51	$5,316.88	$5,520.24	$5,723.60	$5,926.96	$6,130.32	$6,333.68	$6,537.04	$6,740.40	$6,943.76	$7,147.13	$7,350.49
39.00%	$3,205.57	$3,405.65	$3,605.73	$3,805.81	$4,005.89	$4,205.97	$4,406.06	$4,606.14	$4,806.22	$5,006.30	$5,206.38	$5,406.46	$5,606.54	$5,806.62	$6,006.70	$6,206.79	$6,406.87	$6,606.95	$6,807.03	$7,007.11	$7,207.19
40.00%	$3,127.87	$3,324.67	$3,521.48	$3,718.28	$3,915.08	$4,111.88	$4,308.68	$4,505.48	$4,702.28	$4,899.08	$5,095.88	$5,292.68	$5,489.49	$5,686.29	$5,883.09	$6,079.89	$6,276.69	$6,473.49	$6,670.29	$6,867.09	$7,063.89
41.00%	$3,050.18	$3,243.70	$3,437.22	$3,630.74	$3,824.26	$4,017.78	$4,211.30	$4,404.82	$4,598.35	$4,791.87	$4,985.39	$5,178.91	$5,372.43	$5,565.95	$5,759.47	$5,952.99	$6,146.51	$6,340.03	$6,533.56	$6,727.08	$6,920.60
42.00%	$2,972.48	$3,162.72	$3,352.96	$3,543.20	$3,733.45	$3,923.69	$4,113.93	$4,304.17	$4,494.41	$4,684.65	$4,874.89	$5,065.13	$5,255.37	$5,445.61	$5,635.86	$5,826.10	$6,016.34	$6,206.58	$6,396.82	$6,587.06	$6,777.30
43.00%	$2,894.79	$3,081.75	$3,268.71	$3,455.67	$3,642.63	$3,829.59	$4,016.55	$4,203.51	$4,390.47	$4,577.43	$4,764.40	$4,951.36	$5,138.32	$5,325.28	$5,512.24	$5,699.20	$5,886.16	$6,073.12	$6,260.08	$6,447.04	$6,634.01
44.00%	$2,817.09	$3,000.77	$3,184.45	$3,368.13	$3,551.81	$3,735.49	$3,919.18	$4,102.86	$4,286.54	$4,470.22	$4,653.90	$4,837.58	$5,021.26	$5,204.94	$5,388.62	$5,572.30	$5,755.98	$5,939.67	$6,123.35	$6,307.03	$6,490.71
45.00%	$2,739.39	$2,919.79	$3,100.20	$3,280.60	$3,461.00	$3,641.40	$3,821.80	$4,002.20	$4,182.60	$4,363.00	$4,543.40	$4,723.80	$4,904.20	$5,084.61	$5,265.01	$5,445.41	$5,625.81	$5,806.21	$5,986.61	$6,167.01	$6,347.41
46.00%	$2,661.70	$2,838.82	$3,015.94	$3,193.06	$3,370.18	$3,547.30	$3,724.42	$3,901.54	$4,078.66	$4,255.79	$4,432.91	$4,610.03	$4,787.15	$4,964.27	$5,141.39	$5,318.51	$5,495.63	$5,672.75	$5,849.87	$6,027.00	$6,204.12
47.00%	$2,584.00	$2,757.84	$2,931.68	$3,105.52	$3,279.37	$3,453.21	$3,627.05	$3,800.89	$3,974.73	$4,148.57	$4,322.41	$4,496.25	$4,670.09	$4,843.93	$5,017.77	$5,191.62	$5,365.46	$5,539.30	$5,713.14	$5,886.98	$6,060.82
48.00%	$2,506.31	$2,676.87	$2,847.43	$3,017.99	$3,188.55	$3,359.11	$3,529.67	$3,700.23	$3,870.79	$4,041.35	$4,211.91	$4,382.48	$4,553.04	$4,723.60	$4,894.16	$5,064.72	$5,235.28	$5,405.84	$5,576.40	$5,746.96	$5,917.52
49.00%	$2,428.61	$2,595.89	$2,763.17	$2,930.45	$3,097.73	$3,265.01	$3,432.29	$3,599.58	$3,766.86	$3,934.14	$4,101.42	$4,268.70	$4,435.98	$4,603.26	$4,770.54	$4,937.82	$5,105.10	$5,272.38	$5,439.67	$5,606.95	$5,774.23
50.00%	$2,350.91	$2,514.91	$2,678.92	$2,842.92	$3,006.92	$3,170.92	$3,334.92	$3,498.92	$3,662.92	$3,826.92	$3,990.92	$4,154.92	$4,318.92	$4,482.92	$4,646.93	$4,810.93	$4,974.93	$5,138.93	$5,302.93	$5,466.93	$5,630.93

Figure 5.39 – Scenario Analysis Table

5.12 Segmentation Clustering Tool

A final analytical technique of interest is that of segmentation clustering. Figure 5.40 illustrates a sample dataset. You can select the data and run the tool through *Risk Simulator | Tools | Segmentation Clustering*. Figure 5.40 shows a sample segmentation of two groups. That is, taking the original data set, we run some internal algorithms (a combination or k-means hierarchical clustering and other method of moments in order to find the best-fitting groups or natural statistical clusters) to statistically divide, or segment, the original data set into two groups. You can see the two-group memberships in Figure 5.40. Clearly you can segment this data set into as many groups as you wish. This technique is valuable in a variety of settings including marketing (market segmentation of customers into various customer relationship management groups etc.), physical sciences, engineering, and others.



SEGMENTATION AND CLUSTER ANALYSIS RESULT

Sample	Ordered Data	Groups: 2
1	1.00	1
2	1.00	1
3	2.00	1
4	3.00	1
5	2.00	1
6	4.00	1
7	15.00	1
8	16.00	1
9	14.00	1
10	15.00	1
11	125.00	2
12	126.00	2
13	128.00	2
14	129.00	2
15	130.00	2
16	175.00	2
17	179.00	2
18	174.00	2

Figure 5.40 – Segmentation Clustering Tool and Results

5.13 RISK SIMULATOR 2011/2012 NEW TOOLS

5.14 Random Number Generation, Monte Carlo vs. Latin Hypercube, and Correlation Copulas

Starting with version 2011/2012, there are 6 Random Number Generators, 3 Correlation Copulas, and 2 Simulation Sampling Methods to choose from (Figure 5.41). These preferences are set through the *Risk Simulator | Options* location.

The Random Number Generator (RNG) is at the heart of any simulation software. Based on the random number generated, different mathematical distributions can be constructed. The default method is the ROV Risk Simulator proprietary methodology, which provides the best and most robust random numbers. As noted, there are 6 supported random number generators and, in general, the ROV Risk Simulator default method and the Advanced Subtractive Random Shuffle method are the two approaches recommended for use. Do not apply the other methods unless your model or analytics specifically calls for their use, and even then, we recommended testing the results against these two recommended approaches. The further down the list of RNGs, the simpler the algorithm and the faster it runs, in comparison with the more robust results from RNGs further up the list.

In the Correlations section, three methods are supported: the Normal Copula, T-Copula, and Quasi-Normal Copula. These methods rely on mathematical integration techniques, and when in doubt, the normal copula provides the safest and most conservative results. The t-copula provides for extreme values in the tails of the simulated distributions, whereas the quasi-normal copula returns results that are between the values derived by the other two methods.

In the Simulation methods section, Monte Carlo Simulation (MCS) and Latin Hypercube Sampling (LHS) methods are supported. Note that Copulas and other multivariate functions are not compatible with LHS because LHS can be applied to a single random variable but not over a joint distribution. In reality, LHS has very limited impact on the model output's accuracy the more distributions there are in a model since LHS only applies to distributions individually. The benefit of LHS is also eroded if one does not complete the number of samples nominated at the beginning, that is, if one halts the simulation run in mid-simulation. LHS also applies a heavy burden on a simulation model with a large number of inputs because it needs to generate and organize samples from each distribution prior to running the first sample from a distribution. This can cause a long delay in running a large model without providing much more additional accuracy. Finally, LHS is best applied when the distributions are well behaved and symmetrical and without any correlations. Nonetheless, LHS is a powerful approach that yields a uniformly sampled distribution, where MCS can sometimes generate lumpy distributions (sampled data can sometimes be more heavily concentrated in one area of the distribution) as compared to a more uniformly sampled distribution (every part of the distribution will be sampled) when LHS is applied.

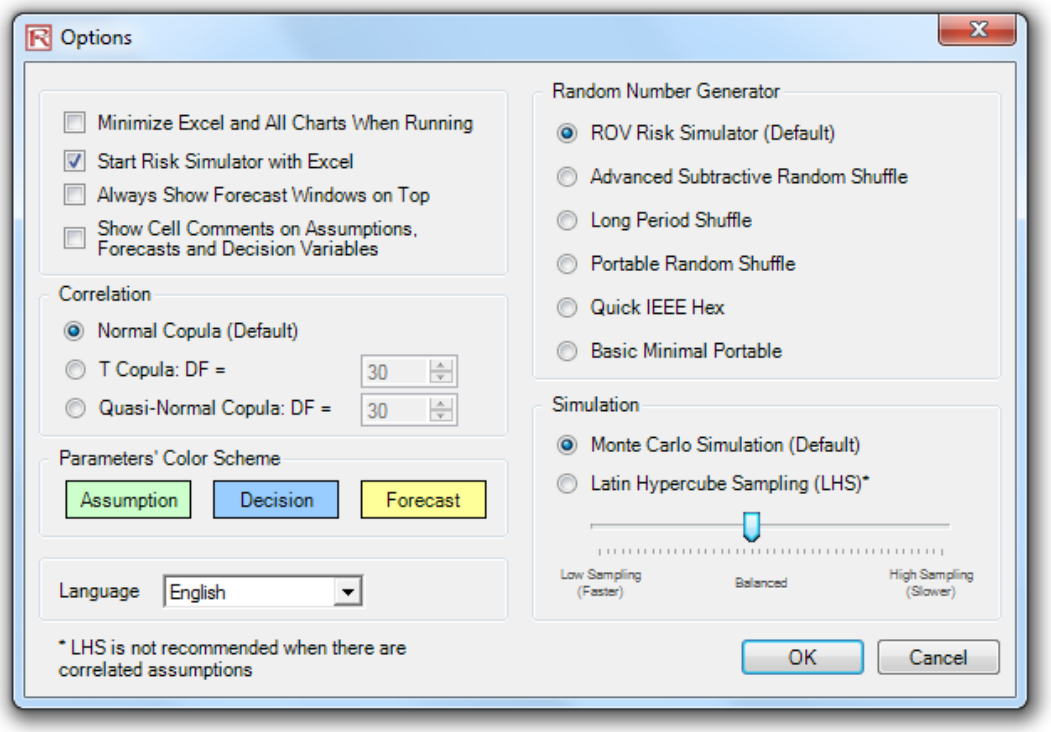


Figure 5.41 – Risk Simulator Options

5.15 Deseasonalizing and Detrending Data

The data deseasonalization and detrending tool removes any seasonal and trending components in your original data (Figure 5.42). In forecasting models, the process usually includes removing the effects of accumulating data sets from seasonality and trend to show only the absolute changes in values and to allow potential cyclical patterns to be identified after removing the general drift, tendency, twists, bends, and effects of seasonal cycles of a set of time-series data. For example, a detrended data set may be necessary to see a more accurate account of a company's sales in a given year more clearly by shifting the entire data set from a slope to a flat surface to better expose the underlying cycles and fluctuations.

Many time-series data exhibit seasonality where certain events repeat themselves after some time period or seasonality period (e.g., ski resorts' revenues are higher in winter than in summer, and this predictable cycle will repeat itself every winter). Seasonality periods represent how many periods would have to pass before the cycle repeats itself (e.g., 24 hours in a day, 12 months in a year, 4 quarters in a year, 60 minutes in an hour, etc.). For deseasonalized and detrended data, a seasonal index greater than 1 indicates a high period or peak within the seasonal cycle, and a value below 1 indicates a dip in the cycle.

Procedure (Deseasonalization and Detrending)

- Select the data you wish to analyze (e.g., *B9:B28*) and click on *Risk Simulator | Tools | Data Deseasonalization and Detrending.*
- Select *Deseasonalize Data* and/or *Detrend Data*, select any detrending models you wish to run, enter in the relevant orders (e.g., polynomial order, moving average order, difference order, and rate order), and click *OK*.
- Review the two reports generated for more details on the methodology, application, and resulting charts and deseasonalized/detrended data.

Procedure (Seasonality Test)

- *Select the data* you wish to analyze (e.g., B9:B28) and click on *Risk Simulator | Tools | Data Seasonality Test.*
- *Enter in the maximum seasonality period to test*. That is, if you enter 6, the tool will test the following seasonality periods: 1, 2, 3, 4, 5, and 6. Period 1, of course, imply no seasonality in the data.
- Review the report generated for more details on the methodology, application, and resulting charts and seasonality test results. The best seasonality periodicity is listed first (ranked by the lowest RMSE error measure), and all the relevant error measurements are included for comparison: root mean squared error (RMSE), mean squared error (MSE), mean absolute deviation (MAD), and mean absolute percentage error (MAPE).

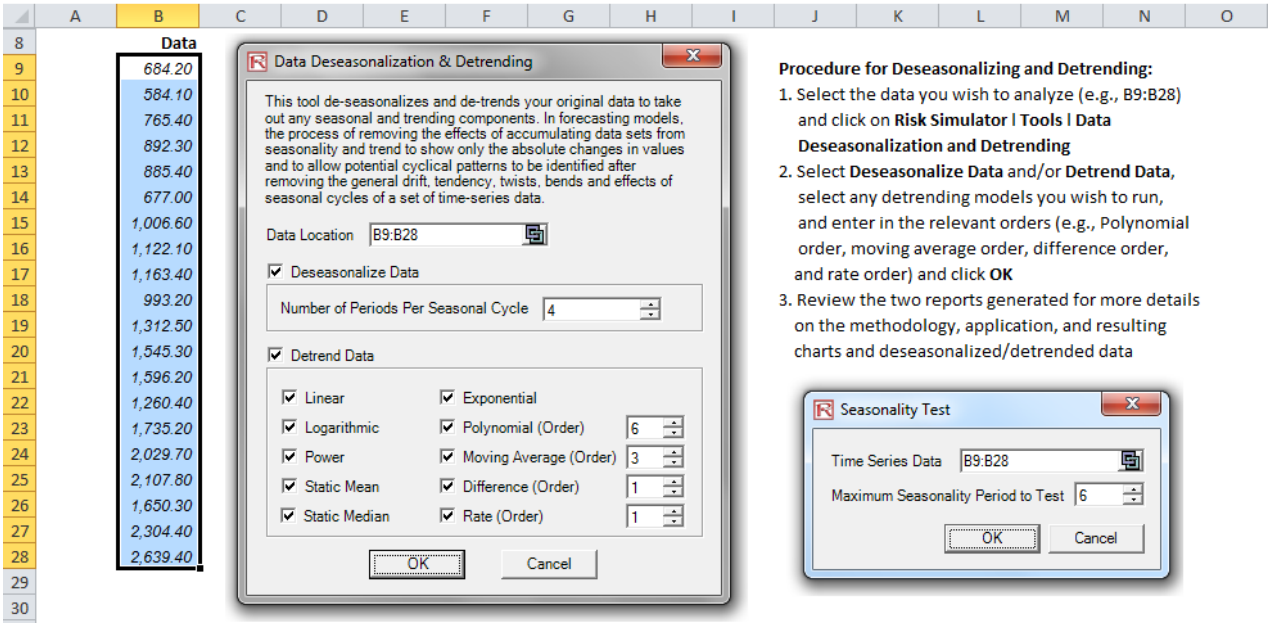


Figure 5.42 – Deseasonalization and Detrending Data

5.16 Principal Component Analysis

Principal Component Analysis is a way of identifying patterns in data and recasting the data in such a way as to highlight their similarities and differences (Figure 5.43). Patterns of data are very difficult to find in high dimensions when multiple variables exist, and higher dimensional graphs are very difficult to represent and interpret. Once the patterns in the data are found, they can be compressed, and the number of dimensions is now reduced. This reduction of data dimensions does not mean much reduction in loss of information. Instead, similar levels of information can now be obtained with a smaller number of variables.

Procedure

- *Select the data* to analyze (e.g., B11:K30), click on *Risk Simulator | Tools | Principal Component Analysis*, and click *OK*.
- Review the generated report for the computed results.

VAR1	VAR2	VAR3	VAR4	VAR5	VAR6	VAR7	VAR8	VAR9	VAR10
96.998	87.223	102.443	112.765	111.984	117.331	78.164	97.658	110.950	89.133
93.098	83.096	81.531	90.224	92.265	78.821	94.321	95.960	101.349	96.345
96.730	96.298	113.426	99.147	98.138	94.868	119.722	108.657	123.757	93.451
116.615	83.876	105.389	109.022	119.189	99.155	94.762	106.751	96.187	107.576
85.558	91.528	84.784	96.371	99.675	100.281	96.773	121.945	82.575	92.635
74.224	114.477	87.202	93.464	107.577	104.667	108.746	105.957	86.282	88.843
106.940	103.226	90.602	97.591	101.315	105.578	101.387	90.890	118.848	104.872
100.722	108.298	108.620	93.635	90.768	111.112	87.988	84.411	107.113	106.384
122.057	114.438	113.039	101.130	100.020	104.537	99.745	89.453	82.252	108.283
104.442	106.179	102.135	89.731	112.382	96.888	91.601	91.789	95.710	95.466
94.762	108.494	105.132	93.917	113.050	82.391	105.506	98.837	100.417	93.459
94.504	108.493	108.030	104.564	106.914	116.306	103.039	105.890	118.528	96.644
110.383	101.435	111.410	98.517	92.202	110.760	94.182	105.339	105.458	96.836
95.592	86.340	119.930	94.335	100.861	97.657	128.354	112.520	108.809	113.322
101.879	105.420	97.504	87.789	112.667	97.111	86.941	107.643	107.843	104.282
104.039	93.519	107.231	105.253	110.750	72.306	104.638	114.671	82.774	100.455
113.540	116.882	102.387	101.451	118.545	99.574	93.431	109.074	99.901	110.392
104.347	114.534	98.788	90.383	84.614	74.349	101.032	102.992	99.822	102.005
102.582	114.762	100.853	88.833	86.101	101.915	109.511	84.912	93.900	105.235
97.832	96.564	98.365	95.603	91.974	106.448	100.588	112.635	102.622	100.571

Procedure:

1. Select the data to analyze (e.g., B11:K30), click on **Risk Simulator | Tools | Principal Component Analysis** and click **OK**
2. Review the generated report for the computed results

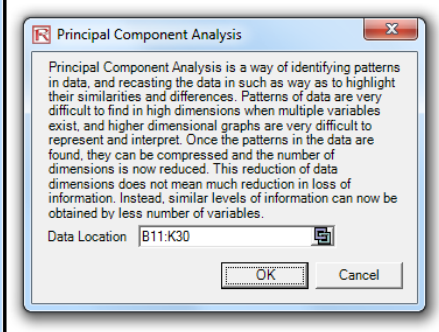


Figure 5.43 – Principal Component Analysis

5.17 Structural Break Analysis

A structural break tests whether the coefficients in different data sets are equal, and this test is most commonly used in time-series analysis to test for the presence of a structural break (Figure 5.44). A time-series data set can be divided into two subsets. Structural break analysis is used to test each subset individually and on one another and on the entire data set to statistically determine if, indeed, there is a break starting at a particular time period. The structural break test is often used to determine whether the independent variables have different impacts on different subgroups of the population, such as to test if a new marketing campaign, activity, major event, acquisition, divestiture, and so forth have an impact on the time-series data.

Suppose, for example, a data set has 100 time-series data points. You can set various breakpoints to test, for instance, data points 10, 30, and 51. (This means that three structural break tests will be performed: data points 1–9 compared with 10–100; data points 1–29 compared with 30–100; and 1–50 compared with 51–100 to see if there is a break in the

underlying structure at the start of data points 10, 30, and 51.). A one-tailed hypothesis test is performed on the null hypothesis (H0) such that the two data subsets are statistically similar to one another, that is, there is no statistically significant structural break. The alternative hypothesis (Ha) is that the two data subsets are statistically different from one another, indicating a possible structural break. If the calculated p-values are less than or equal to 0.01, 0.05, or 0.10, then the hypothesis is rejected, which implies that the two data subsets are statistically significantly different at the 1%, 5%, and 10% significance levels. High p-values indicate that there is no statistically significant structural break.

Procedure

- *Select the data* you wish to analyze (e.g., B15:D34), click on *Risk Simulator | Tools | Structural Break Test*, enter in the relevant test points you wish to apply on the data (e.g., 6, 10, 12), and click *OK*.
- Review the report to determine which of these test points indicate a statistically significant break point in your data and which points do not.

Y	X1	X2
521	18308	185
367	1148	600
443	18068	372
365	7729	142
614	100484	432
385	16728	290
286	14630	346
397	4008	328
764	38927	354
427	22322	266
153	3711	320
231	3136	197
524	50508	266
328	28886	173
240	16996	190
286	13035	239
285	12973	190
569	16309	241
96	5227	189
498	19235	358

Procedure:

1. Select the data to analyze (e.g., B15:D34) and click on **Risk Simulator | Tools | Structural Break Test**, and enter in the relevant test points you wish to apply on the data (e.g., 6, 10, 12) and click **OK**
2. Review the report to determine which of these test points indicate a statistically significant break point in your data and which points do not

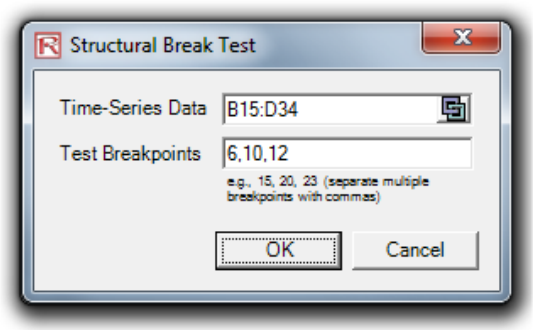


Figure 5.44 – Structural Break Analysis

5.18 Trendline Forecasts

Trendlines can be used to determine if a set of time-series data follows any appreciable trend (Figure 5.45). Trends can be linear or nonlinear (such as exponential, logarithmic, moving average, power, polynomial, or power).

Procedure

- *Select the data* you wish to analyze, click on *Risk Simulator | Forecasting | Trendline*, select the relevant trendlines you wish to apply on the data (e.g., select all methods by default), enter in the number of periods to forecast (e.g., 6 periods), and click OK.
- Review the report to determine which of these test trendlines provide the best fit and best forecast for your data.

Historical Sales Revenues

Year	Quarter	Period	Sales
2006	1	1	$684.20
2006	2	2	$584.10
2006	3	3	$765.40
2006	4	4	$892.30
2007	1	5	$885.40
2007	2	6	$677.00
2007	3	7	$1,006.60
2007	4	8	$1,122.10
2008	1	9	$1,163.40
2008	2	10	$993.20
2008	3	11	$1,312.50
2008	4	12	$1,545.30
2009	1	13	$1,596.20
2009	2	14	$1,260.40
2009	3	15	$1,735.20
2009	4	16	$2,029.70
2010	1	17	$2,107.80
2010	2	18	$1,650.30
2010	3	19	$2,304.40
2010	4	20	$2,639.40

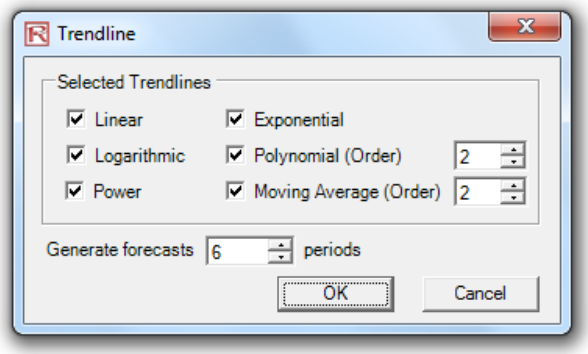


Figure 5.45 – Trendline Forecasts

5.19 Model Checking Tool

After a model is created and after assumptions and forecasts have been set, you can run the simulation as usual or run the Check Model tool (Figure 5.46) to test if the model has been set up correctly. Alternatively, if the model does not run and you suspect that some settings may be incorrect, run this tool from *Risk Simulator | Tools | Check Model* to identify where there might be problems with your model. Note that while this tool checks for the most common model problems as well as for problems in Risk Simulator assumptions and forecasts, it is in no way comprehensive enough to test for all types of problems. It is still up to the model developer to make sure the model works properly.

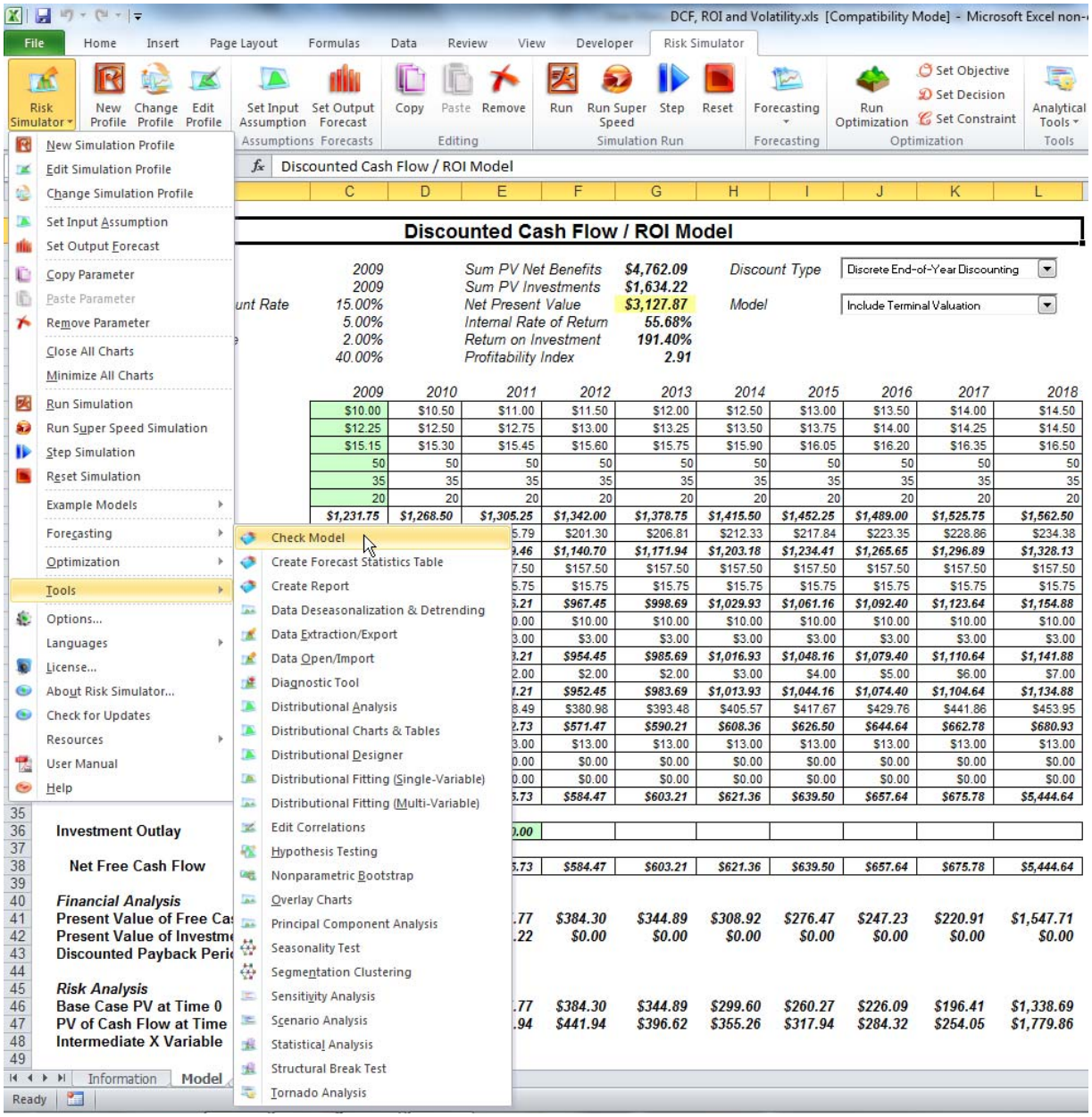


Figure 5.46 – Model Checking Tool

5.20 Percentile Distributional Fitting Tool

The Percentile Distributional Fitting tool (Figure 5.47) is another alternate way of fitting probability distributions. There are several related tools and each has its own uses and advantages:

- **Distributional Fitting (Percentiles)**—using an alternate method of entry (percentiles and first/second moment combinations) to find the best-fitting parameters of a specified distribution without the need for having raw data. This

method is suitable for use when there are insufficient data, only when percentiles and moments are available, or as a means to recover the entire distribution with only two or three data points but the distribution type needs to be assumed or known.

- **Distributional Fitting (Single Variable)**—using statistical methods to fit your raw data to all 42 distributions to find the best fitting distribution and its input parameters. Multiple data points are required for a good fit, and the distribution type may or may not be known ahead of time.
- **Distributional Fitting (Multiple Variables)**—using statistical methods to fit your raw data on multiple variables at the same time. This method uses the same algorithms as the single variable fitting, but incorporates a pairwise correlation matrix between the variables. Multiple data points are required for a good fit, and the distribution type may or may not be known ahead of time.
- **Custom Distribution (Set Assumption)**—using nonparametric resampling techniques to generate a custom distribution with the existing raw data and to simulate the distribution based on this empirical distribution. Fewer data points are required, and the distribution type is not known ahead of time.

Procedure

Click on *Risk Simulator | Tools | Distributional Fitting (Percentiles)*, choose the probability distribution and types of inputs you wish to use, enter the parameters, and click Run to obtain the results. Review the fitted R-square results and compare the empirical versus theoretical fitting results to determine if your distribution is a good fit.

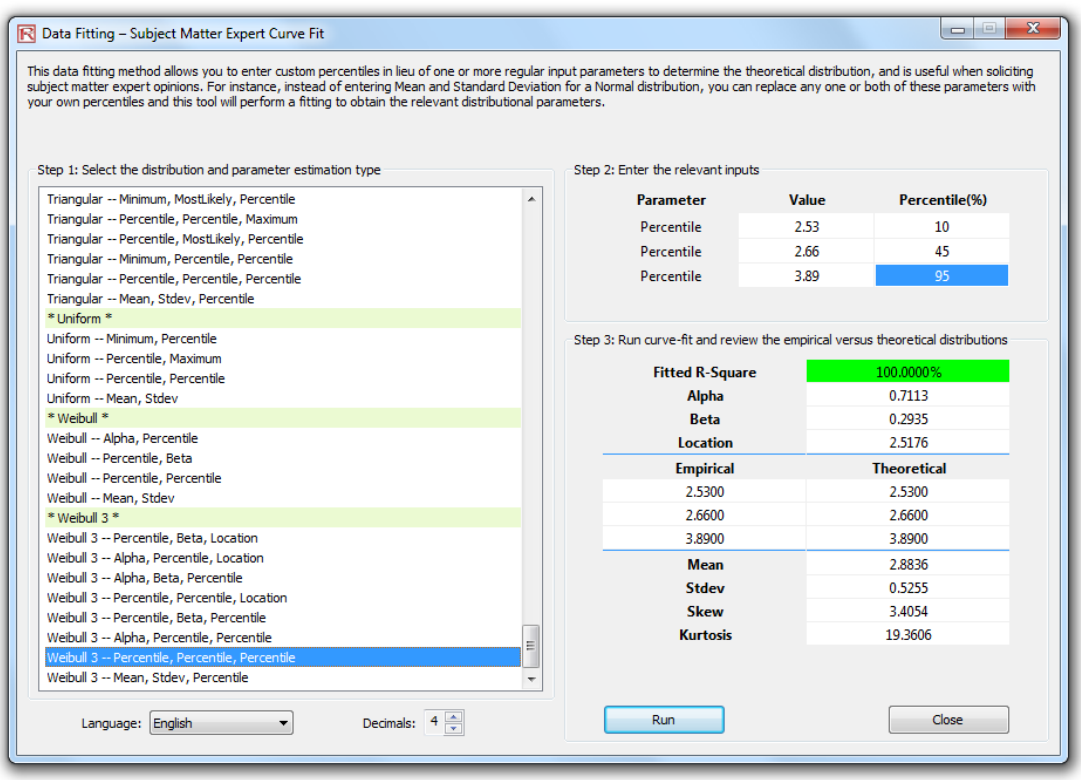


Figure 5.47 – Percentile Distributional Fitting Tool

5.21 Distribution Charts and Tables: Probability Distribution Tool

Distributional Charts and Tables is a new Probability Distribution tool that is a very powerful and fast module used for generating distribution charts and tables (Figures 5.48 through 5.51). Note that there are three similar tools in Risk Simulator but each does very different things.

Distributional Analysis

Used to quickly compute the PDF, CDF, and ICDF of the 42 probability distributions available in Risk Simulator, and to return a probability table of these values.

Distributional Charts and Tables

The Probability Distribution tool described here used to compare different parameters of the same distribution (e.g., the shapes and PDF, CDF, ICDF values of a Weibull distribution with Alpha and Beta of [2, 2], [3, 5], and [3.5, 8], and overlays them on top of one another).

Overlay Charts

Used to compare different distributions (theoretical input assumptions and empirically simulated output forecasts) and to overlay them on top of one another for a visual comparison.

Procedure

- Run ROV BizStats at *Risk Simulator | Distributional Charts and Tables*, click on the Apply Global Inputs button to load a sample set of input parameters or enter your own inputs, and click Run to compute the results. The resulting four moments and CDF, ICDF, PDF are computed for each of the 45 probability distributions (Figure 5.48).
- Click on the Charts and Tables tab (Figure 5.49), select a distribution [A] (e.g., Arcsine), choose if you wish to run the CDF, ICDF, or PDF [B], enter the relevant inputs, and click Run Chart or Run Table [C]. You can switch between the Charts and Table tab to view the results as well as try out some of the chart icons [E] to see the effects on the chart.
- You can also change two parameters [H] to generate multiple charts and distribution tables by entering the From/To/Step input or using the Custom inputs and then hitting Run. For example, as illustrated in Figure 5.50, run the Beta distribution and select PDF [G], select Alpha and Beta to change [H] using custom [I] inputs and enter the relevant input parameters: 2;5;5 for Alpha and 5;3;5 for Beta [J], and click Run Chart. This will generate three Beta distributions [K]: Beta (2,5), Beta (5,3), and Beta (5,5) [L]. Explore various chart types, gridlines, language, and decimal settings [M], and try rerunning the distribution using theoretical versus empirically simulated values [N].
- Figure 5.51 illustrates the probability tables generated for a binomial distribution where the probability of success and number of successful trials (random variable X) are selected to vary [O] using the From/To/Step option. Try to replicate the calculation as shown and click on the Table tab [P] to view the created probability density function results. This example uses a binomial distribution with a starting input set of Trials = 20, Probability (of success) = 0.5, and Random X, or Number of Successful Trials, = 10, where the Probability of Success is allowed to change from 0., 0.25, …, 0.50 and is shown as the row variable, and the Number of Successful Trials is also allowed to change from 0, 1, 2, …, 8, and is shown as the column variable. PDF is chosen and, hence, the results in the table show the probability that the given events occur. For instance, the probability of getting exactly 2 successes when 20 trials are run where each trial has a 25% chance of success is 0.0669, or 6.69%.

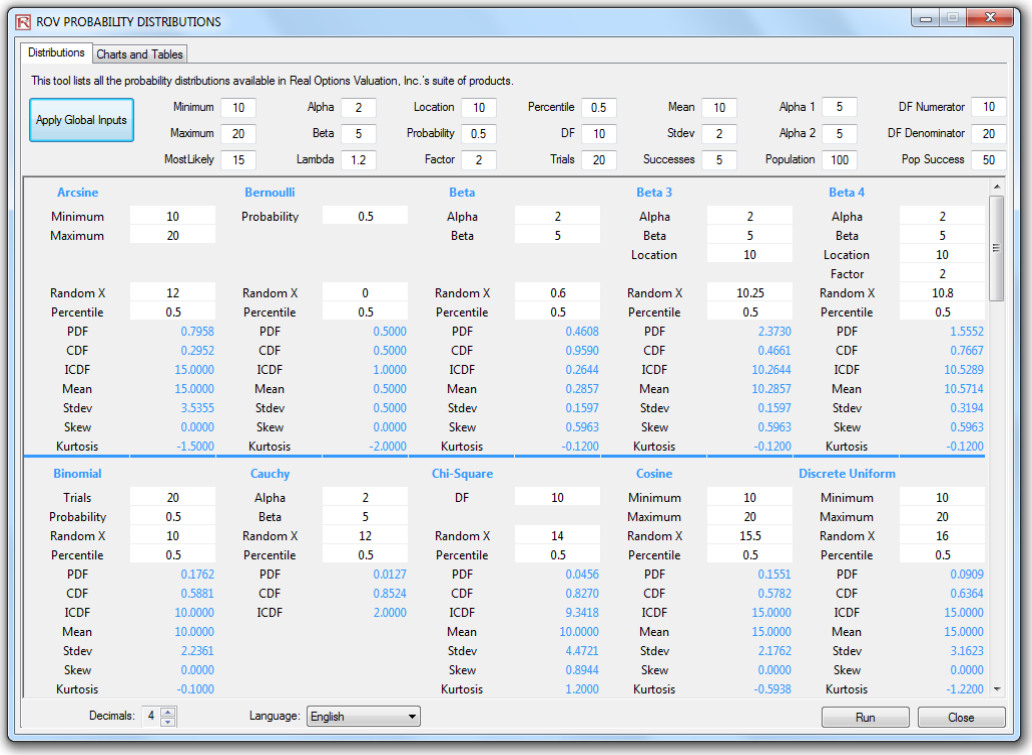


Figure 5.48 – Probability Distribution Tool (45 Probability Distributions)

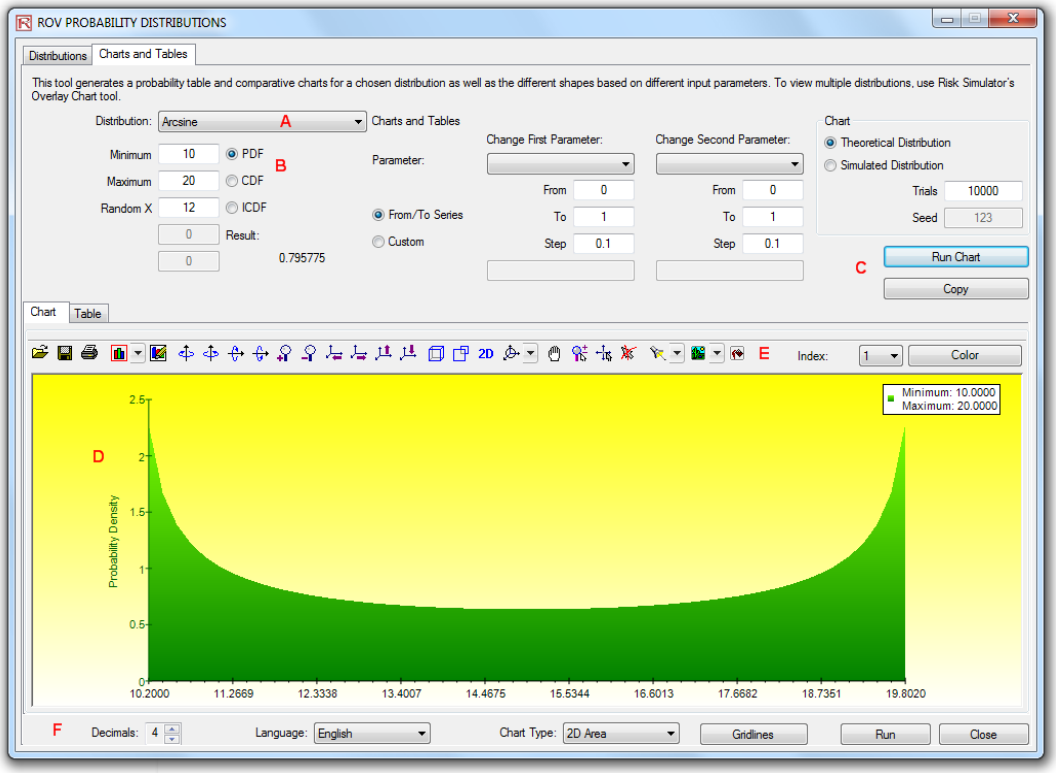


Figure 5.49 – ROV Probability Distribution (PDF and CDF Charts)

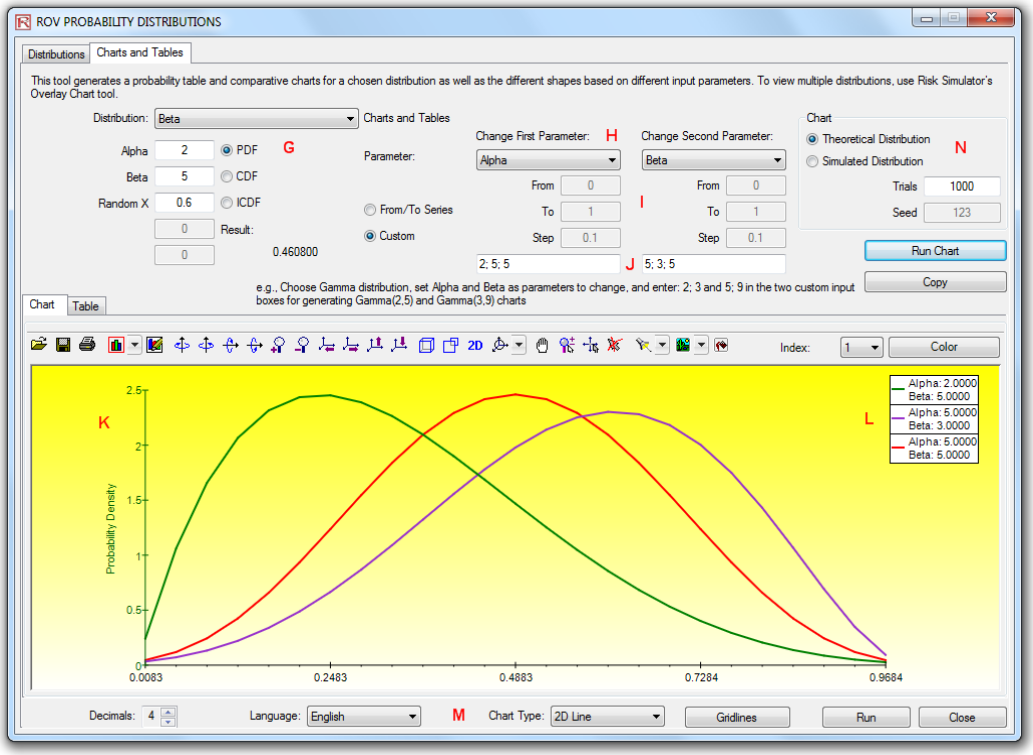
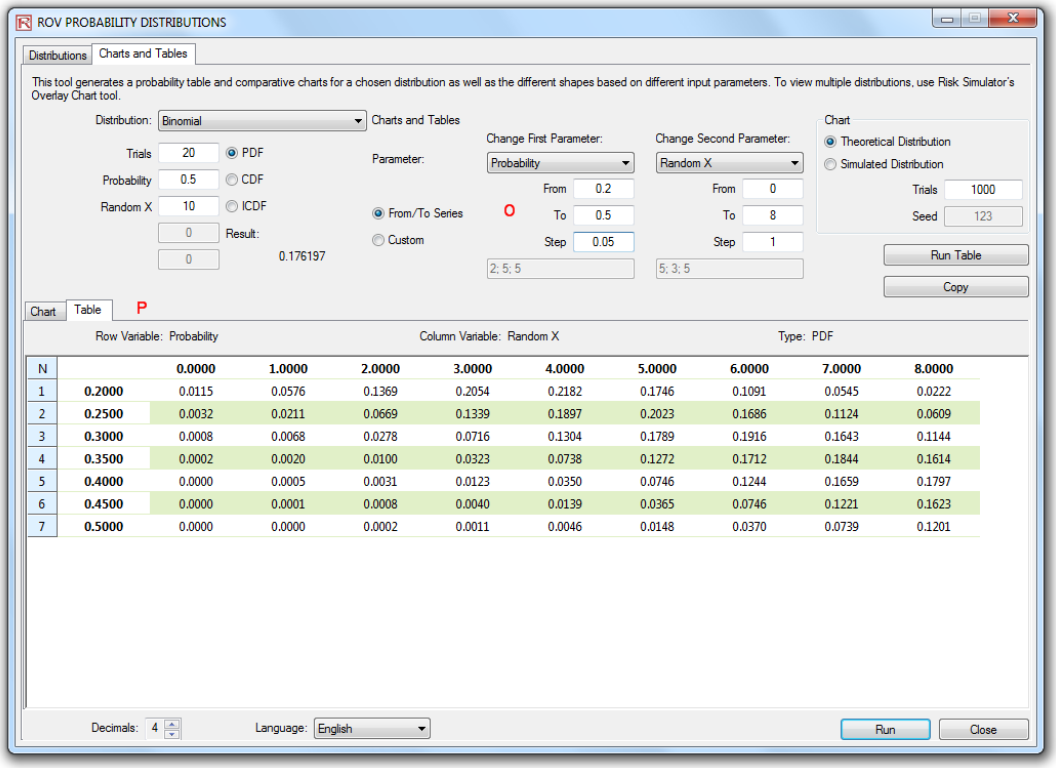


Figure 5.50 – ROV Probability Distribution (Multiple Overlay Charts)



N		0.0000	1.0000	2.0000	3.0000	4.0000	5.0000	6.0000	7.0000	8.0000
1	**0.2000**	0.0115	0.0576	0.1369	0.2054	0.2182	0.1746	0.1091	0.0545	0.0222
2	**0.2500**	0.0032	0.0211	0.0669	0.1339	0.1897	0.2023	0.1686	0.1124	0.0609
3	**0.3000**	0.0008	0.0068	0.0278	0.0716	0.1304	0.1789	0.1916	0.1643	0.1144
4	**0.3500**	0.0002	0.0020	0.0100	0.0323	0.0738	0.1272	0.1712	0.1844	0.1614
5	**0.4000**	0.0000	0.0005	0.0031	0.0123	0.0350	0.0746	0.1244	0.1659	0.1797
6	**0.4500**	0.0000	0.0001	0.0008	0.0040	0.0139	0.0365	0.0746	0.1221	0.1623
7	**0.5000**	0.0000	0.0000	0.0002	0.0011	0.0046	0.0148	0.0370	0.0739	0.1201

Figure 5.51 – ROV Probability Distribution (Distribution Tables)

5.22 ROV BizStats

This new ROV BizStats tool is a very powerful and fast module in Risk Simulator that is used for running business statistics and analytical models on your data. It covers more than 130 business statistics and analytical models (Figures 5.52 through 5.55). The following provides a few quick getting started steps on running the module and details on each of the elements in the software.

Procedure

- Run ROV BizStats at *Risk Simulator | ROV BizStats* and click on Example to load a sample data and model profile [A] or type in your data or copy/paste into the data grid [D] (Figure 5.52). You can add your own notes or variable names in the first Notes row [C].
- Select the relevant model [F] to run in Step 2 and using the example data input settings [G], enter in the relevant variables [H]. Separate variables for the same parameter using semicolons and use a new line (hit Enter to create a new line) for different parameters.
- Click Run [I] to compute the results [J]. You can view any relevant analytical results, charts, or statistics from the various tabs in Step 3.
- If required, you can provide a model name to save into the profile in Step 4 [L]. Multiple models can be saved in the same profile. Existing models can be edited or deleted [M] and rearranged in order of appearance [N], and all the changes can be saved [O] into a single profile with the file name extension *.bizstats.

Notes

- The data grid size can be set in the menu, where the grid can accommodate up to 1,000 variable columns with 1 million rows of data per variable. The menu also allows you to change the language settings and decimal settings for your data.
- To get started, it is always a good idea to load the example file [A] that comes complete with some data and precreated models [S]. You can double-click on any of these models to run them and the results are shown in the report area [J], which sometimes can be a chart or model statistics [T/U]. Using this example file, you can now see how the input parameters [H] are entered based on the model description [G], and you can proceed to create your own custom models.
- Click on the variable headers [D] to select one or multiple variables at once, and then right-click to add, delete, copy, paste, or visualize [P] the variables selected.
- Models can also be entered using a Command console [V/W/X]. To see how this works, double-click to run a model [S] and go to the Command console [V]. You can replicate the model or create your own and click Run Command [X] when ready. Each line in the console represents a model and its relevant parameters.
- The entire *.bizstats profile (where data and multiple models are created and saved) can be edited directly in XML [Z] by opening the XML Editor from the File menu. Changes to the profile can be programmatically made here and takes effect once the file is saved.
- Click on the data grid's column header(s) to select the entire column(s) or variable(s), and once selected, you can right-click on the header to Auto Fit the column, Cut,

Copy, Delete, or Paste data. You can also click on and select multiple column headers to select multiple variables and right-click and select Visualize to chart the data.

- If a cell has a large value that is not completely displayed, click on and hover your mouse over that cell and you will see a popup comment showing the entire value, or simply resize the variable column (drag the column to make it wider, double click on the column's edge to auto fit the column, or right click on the column header and select auto fit).
- Use the up, down, left, right keys to move around the grid, or use the Home and End keys on the keyboard to move to the far left and far right of a row. You can also use combination keys such as: Ctrl+Home to jump to the top left cell, Ctrl+End to the bottom right cell, Shift+Up/Down to select a specific area, and so forth.
- You can enter short notes for each variable on the Notes row. Remember to make your notes short and simple.
- Try out the various chart icons on the Visualize tab to change the look and feel of the charts (e.g., rotate, shift, zoom, change colors, add legend, and so forth).
- The Copy button is used to copy the Results, Charts, and Statistics tabs in Step 3 after a model is run. If no models are run, then the copy function will only copy a blank page.
- The Report button will only run if there are saved models in Step 4 or if there is data in the grid, else the report generated will be empty. You will also need Microsoft Excel to be installed to run the data extraction and results reports, and Microsoft PowerPoint available to run the chart reports.
- When in doubt about how to run a specific model or statistical method, start the Example profile and review how the data is setup in Step 1 or how the input parameters are entered in Step 2. You can use these as getting started guides and templates for your own data and models.
- The language can be changed in the Language menu. Note that currently there are 10 languages available in the software with more to be added later. However, sometimes certain limited results will still be shown in English.
- You can change how the list of models in Step 2 is shown by changing the View drop list. You can list the models alphabetically, categorically, and by data input requirements—note that in certain Unicode languages (e.g., Chinese, Japanese, and Korean), there is no alphabetical arrangement and therefore the first option will be unavailable.
- The software can handle different regional decimal and numerical settings (e.g., one thousand dollars and fifty cents can be written as 1,000.50 or 1.000,50 or 1'000,50 and so forth). The decimal settings can be set in ROV BizStats' menu *Data | Decimal Settings*. However, when in doubt, please change the computer's regional settings to English USA and keep the default North America 1,000.50 in ROV BizStats (this setting is guaranteed to work with ROV BizStats and the default examples).

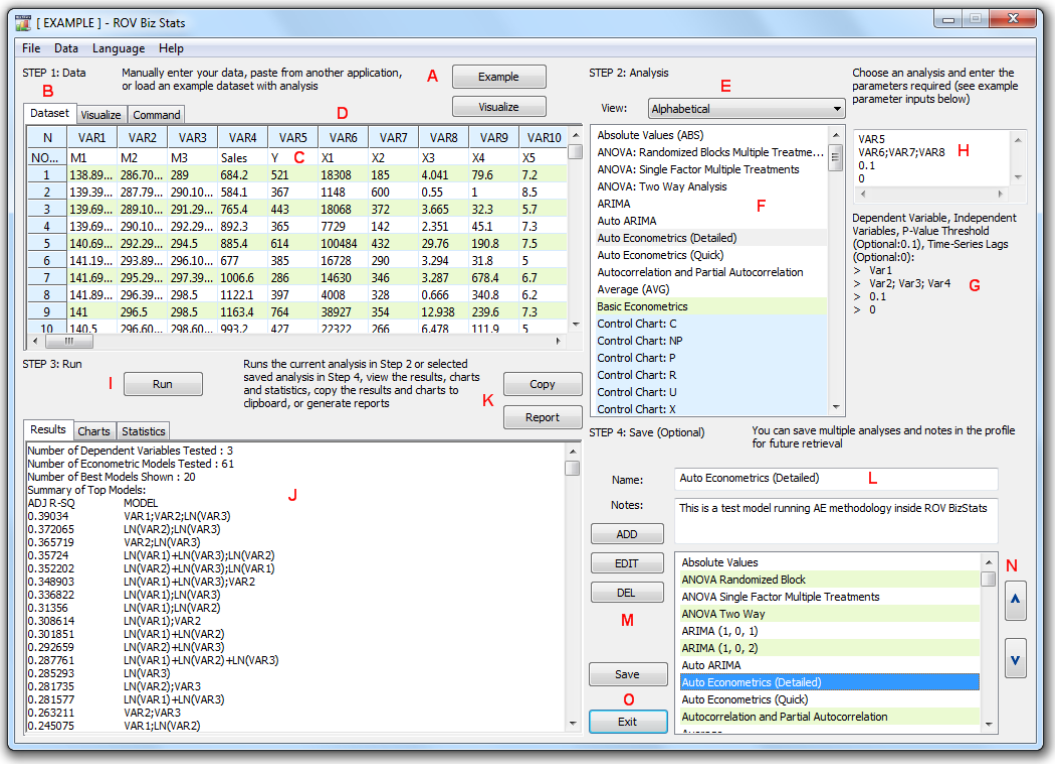


Figure 5.52 – ROV BizStats (Statistical Analysis)

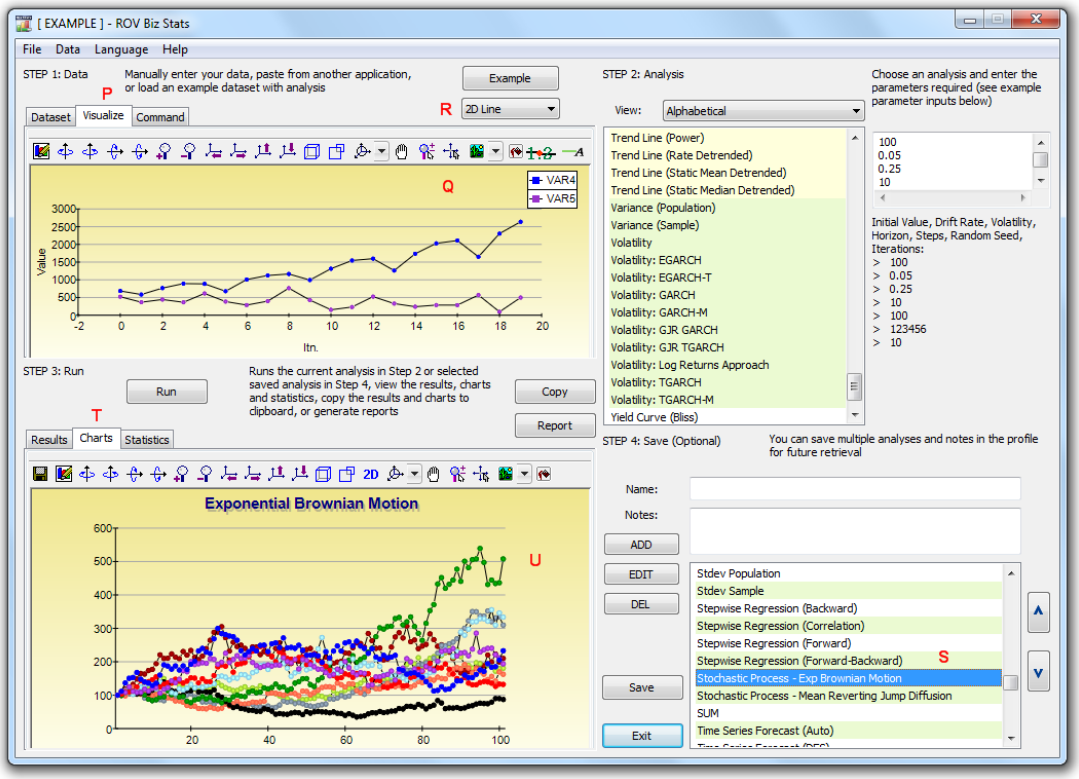


Figure 5.53 – ROV BizStats (Data Visualization and Results Charts)

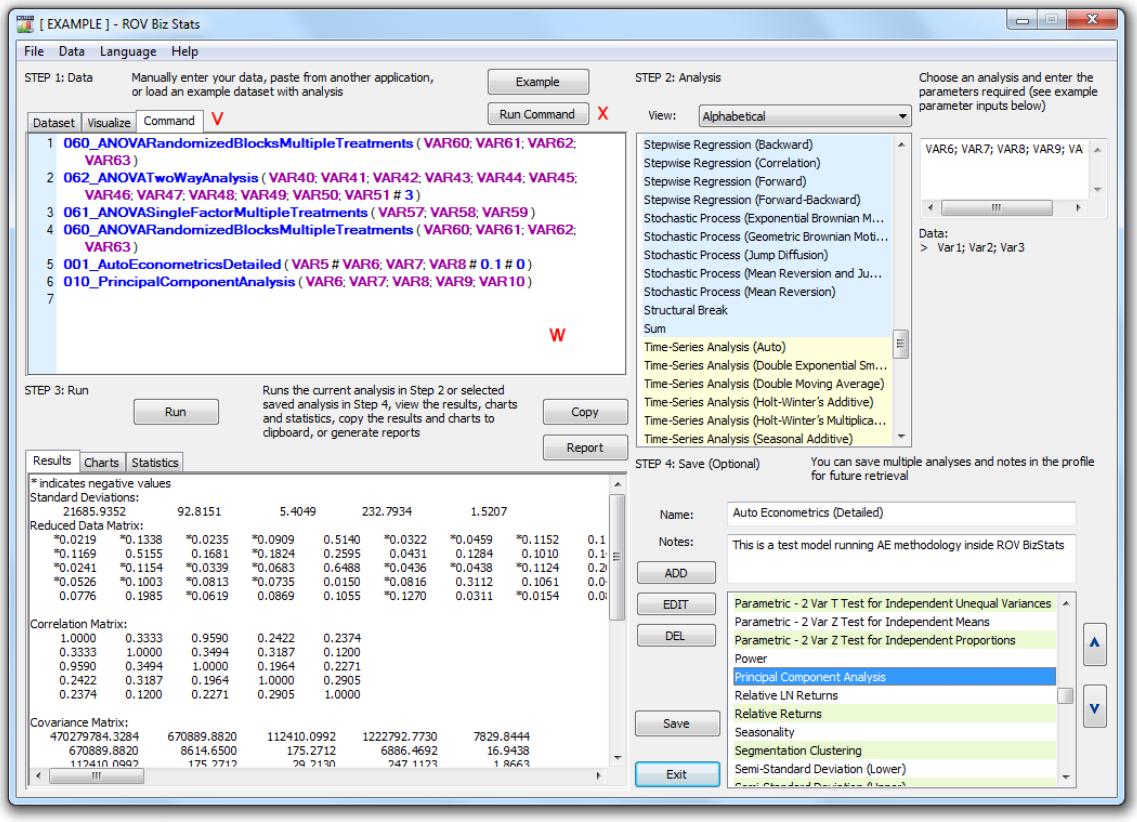


Figure 5.54 – ROV BizStats (Command Console)

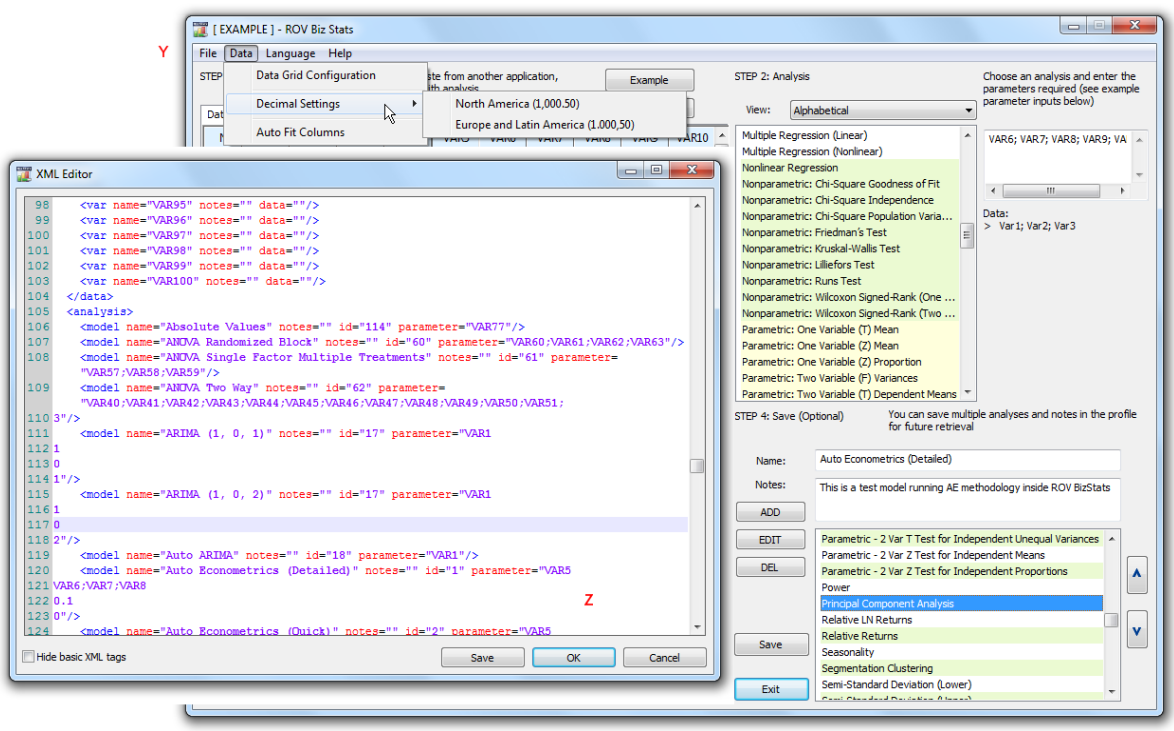


Figure 5.55 – ROV BizStats (XML Editor)

5.23 Neural Network and Combinatorial Fuzzy Logic Forecasting Methodologies

The term Neural Network is often used to refer to a network or circuit of biological neurons, while modern usage of the term often refers to artificial neural networks comprising artificial neurons, or nodes, recreated in a software environment. Such networks attempt to mimic the neurons in the human brain in ways of thinking and identifying patterns and, in our situation, identifying patterns for the purposes of forecasting time-series data. In Risk Simulator, the methodology is found inside the ROV BizStats module located at *Risk Simulator | ROV BizStats | Neural Network* as well as in *Risk Simulator | Forecasting | Neural Network*. Figure 5.56 shows the Neural Network forecast methodology.

Procedure

- Click on *Risk Simulator | Forecasting | Neural Network*.
- Start by either manually entering data or pasting some data from the clipboard (e.g., select and copy some data from Excel, start this tool, and paste the data by clicking on the Paste button).
- Select if you wish to run a *Linear* or *Nonlinear Neural Network* model, enter in the desired number of Forecast Periods (e.g., 5), the number of hidden Layers in the Neural Network (e.g., 3), and number of Testing Periods (e.g., 5).
- Click *Run* to execute the analysis and review the computed results and charts. You can also *Copy* the results and chart to the clipboard and paste it in another software application.

Note that the number of hidden layers in the network is an input parameter and will need to be calibrated with your data. Typically, the more complicated the data pattern, the higher the number of hidden layers you would need and the longer it would take to compute. It is recommended that you start at 3 layers. The testing period is simply the number of data points used in the final calibration of the Neural Network model, and we recommend using at least the same number of periods you wish to forecast as the testing period.

In contrast, the term fuzzy logic is derived from fuzzy set theory to deal with reasoning that is approximate rather than accurate—as opposed to crisp logic, where binary sets have binary logic, fuzzy logic variables may have a truth value that ranges between 0 and 1 and is not constrained to the two truth values of classic propositional logic. This fuzzy weighting schema is used together with a combinatorial method to yield time-series forecast results in Risk Simulator as illustrated in Figure 5.57, and is most applicable when applied to time-series data that has seasonality and trend. This methodology is found inside the ROV BizStats module in Risk Simulator, at *Risk Simulator | ROV BizStats | Combinatorial Fuzzy Logic* as well as *in Risk Simulator | Forecasting | Combinatorial Fuzzy Logic*.

Procedure

- Click on *Risk Simulator | Forecasting | Combinatorial Fuzzy Logic*.
- Start by either manually entering data or pasting some data from the clipboard (e.g., select and copy some data from Excel, start this tool, and paste the data by clicking on the Paste button)

- Select the variable you wish to run the analysis on from the drop-down list, and enter in the seasonality period (e.g., 4 for quarterly data, 12 for monthly data, etc.) and the desired number of Forecast Periods (e.g., 5).
- Click *Run* to execute the analysis and review the computed results and charts. You can also Copy the results and chart to the clipboard and paste it in another software application.

Note that neither neural networks nor fuzzy logic techniques have yet been established as valid and reliable methods in the business forecasting domain, on either a strategic, tactical, or operational level. Much research is still required in these advanced forecasting fields. Nonetheless, Risk Simulator provides the fundamentals of these two techniques for the purposes of running time-series forecasts. We recommend that you do not use any of these techniques in isolation, but, rather, in combination with the other Risk Simulator forecasting methodologies to build more robust models.

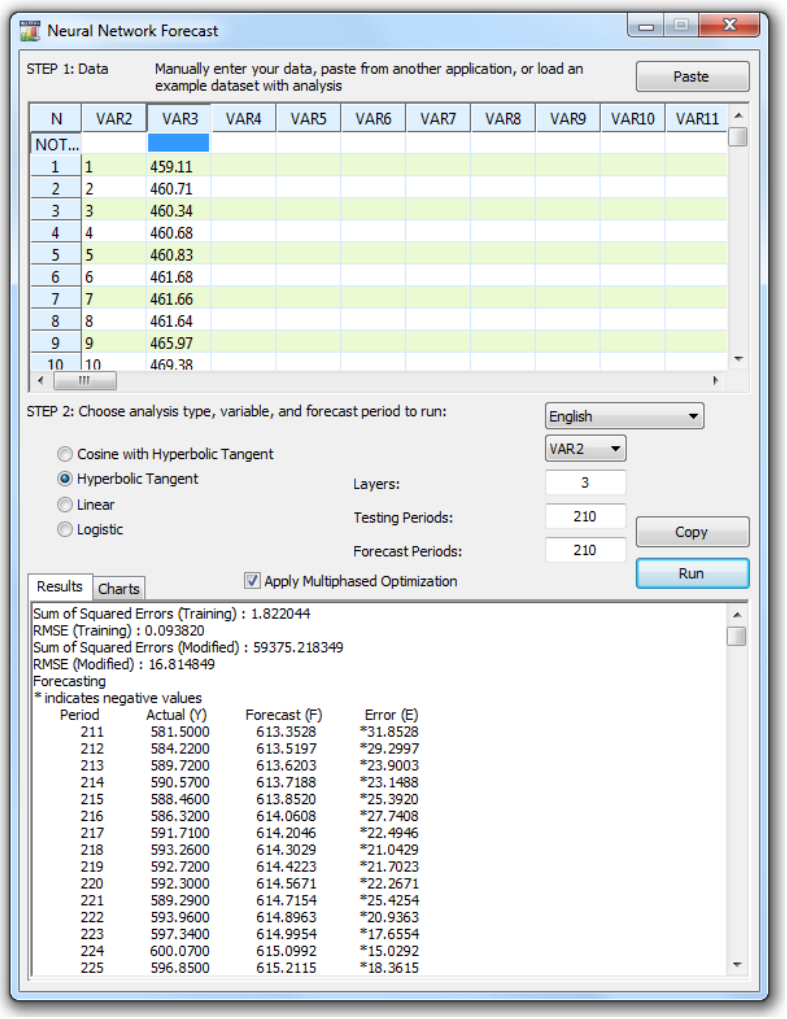


Figure 5.56 – Neural Network Forecast

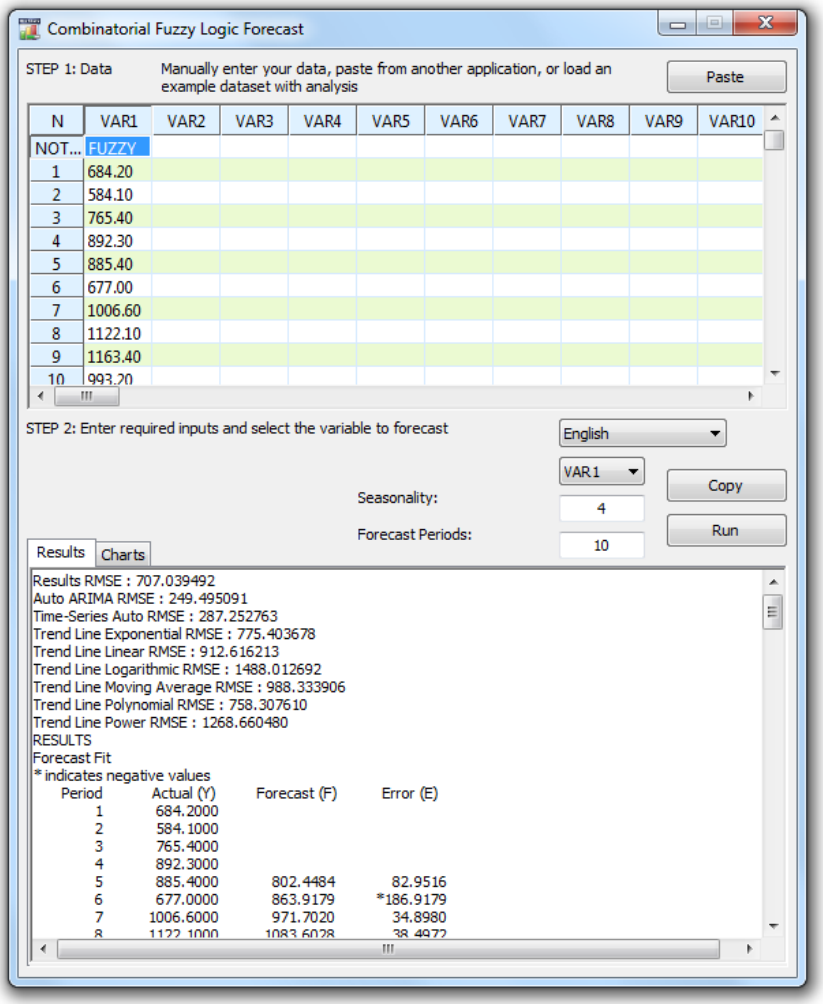


Figure 5.57 – Fuzzy Logic Time-Series Forecast

5.24 Optimizer Goal Seek

The Goal Seek tool is a search algorithm applied to find the solution of a single variable within a model. If you know the result that you want from a formula or a model, but are not sure what input value the formula needs to get that result, use the *Risk Simulator | Tools | Goal Seek* feature. Note that Goal Seek works only with one variable input value. If you want to accept more than one input value, use Risk Simulator's advanced Optimization routines. Figure 5.58 shows how Goal Seek is applied to a simple model.

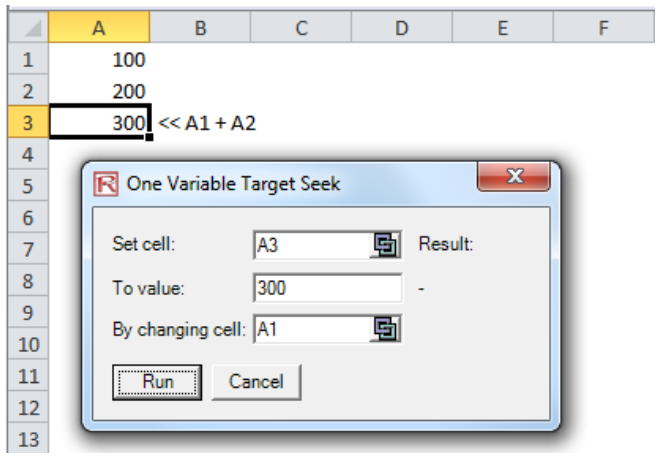


Figure 5.58 – Goal Seek

5.25 Single Variable Optimizer

The Single Variable Optimizer tool is a search algorithm used to find the solution of a single variable within a model, just like the goal seek routine discussed previously. If you want the maximum or minimum possible result from a model but are not sure what input value the formula needs to get that result, use the *Risk Simulator | Tools | Single Variable Optimizer* feature (Figure 5.59). Note that this tool runs very quickly but is only applicable to finding one variable input. If you want to accept more than one input value, use Risk Simulator's advanced Optimization routines. Note that this tool is included in Risk Simulator because if you require a quick optimization computation for a single decision variable, this tool provides that capability without having to set up an optimization model with profiles, simulation assumptions, decision variables, objectives, and constraints.

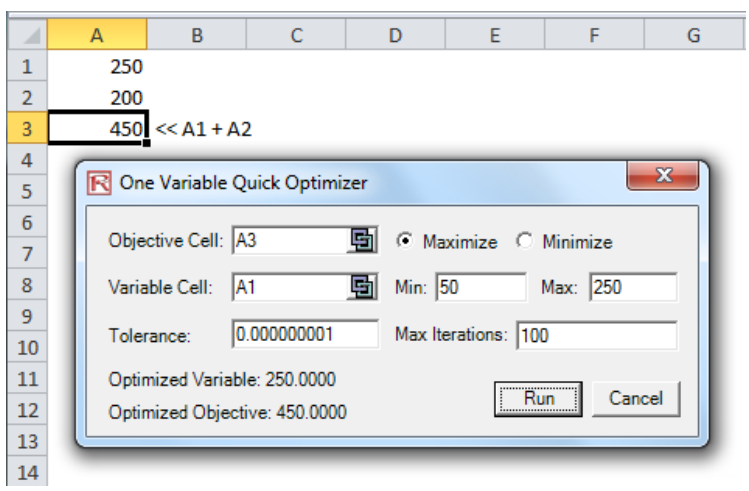


Figure 5.59 – Single Variable Optimizer

5.26 Genetic Algorithm Optimization

Genetic Algorithms belong to the larger class of evolutionary algorithms that generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. Genetic Algorithm is a search heuristic that mimics the process of natural evolution and is routinely used to generate useful solutions to optimization and search problems.

The genetic algorithm is available in *Risk Simulator | Tools | Genetic Algorithm* (Figure 5.60). Care should be taken in calibrating the model's inputs as the results will be fairly sensitive to the inputs (the default inputs are provided as a general guide to the most common input levels), and it is recommended that the Gradient Search Test option be chosen for a more robust set of results (you can deselect this option to get started and then select this choice, rerun the analysis, and compare the results).

Notes

In many problems, genetic algorithms may have a tendency to converge towards local optima or even arbitrary points rather than the global optimum of the problem. This means that it does not know how to sacrifice short-term fitness to gain longer-term fitness. For specific optimization problems and problem instances, other optimization algorithms may find better solutions than genetic algorithms (given the same amount of computation time). Therefore, it is recommended that you first run the Genetic Algorithm and then rerun it by selecting the Apply Gradient Search Test option (Figure 5.60) to check the robustness of the model. This gradient search test will attempt to run combinations of traditional optimization techniques with Genetic Algorithm methods and return the best possible solution. Finally, unless there is a specific theoretical need to use Genetic Algorithm, we recommend using Risk Simulator's Optimization module, which allows you to run more advanced risk-based dynamic and stochastic optimization routines, for more robust results.

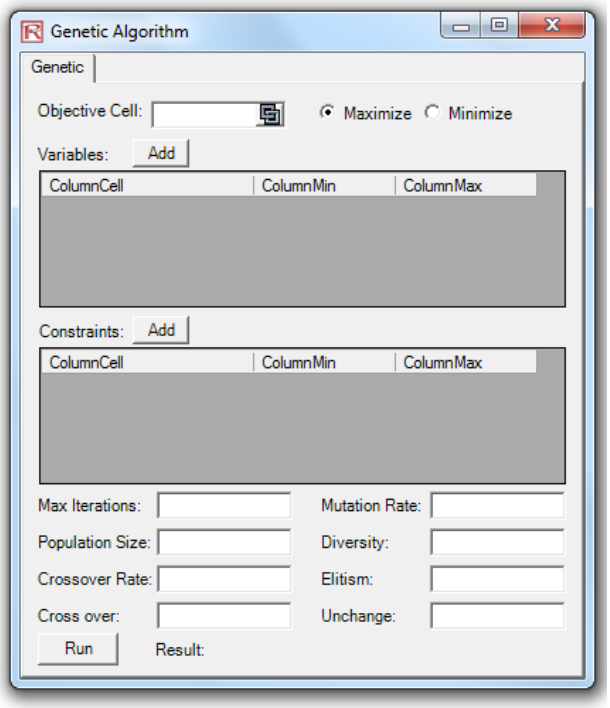


Figure 5.60 – Genetic Algorithm

5.27 ROV Decision Tree Module

5.27.1 Decision Tree

Risk Simulator | ROV Decision Tree runs the Decision Tree module (Figure 5.61). ROV Decision Tree is used to create and value decision tree models. Additional advanced methodologies and analytics are also included:

- Decision Tree Models
- Monte Carlo risk simulation
- Sensitivity Analysis
- Scenario Analysis
- Bayesian (Joint and Posterior Probability Updating)
- Expected Value of Information
- MINIMAX
- MAXIMIN
- Risk Profiles

The following are some main quick getting started tips and procedures for using this intuitive tool:

- There are 11 localized languages available in this module and the current language can be changed through the Language menu.
- Insert Option nodes or Insert Terminal nodes by first selecting any existing node and then clicking on the option node icon (square) or terminal node icon (triangle), or use the functions in the Insert menu.
- Modify individual Option Node or Terminal Node properties by double-clicking on a node. Sometimes when you click on a node, all subsequent child nodes are also selected (this allows you to move the entire tree starting from that selected node). If you wish to select only that node, you may have to click on the empty background and click back on that node to select it individually. Also, you can move individual nodes or the entire tree started from the selected node depending on the current setting (right-click, or in the Edit menu, and select Move Nodes Individually or Move Nodes Together).
- The following are some quick descriptions of the things that can be customized and configured in the node properties user interface. It is simplest to try different settings for each of the following to see its effects in the Strategy Tree:
    - Name. Name shown above the node.
    - Value. Value shown below the node.
    - Excel Link. Links the value from an Excel spreadsheet's cell.
    - Notes. Notes can be inserted above or below a node.
    - Show in Model. Show any combinations of Name, Value, and Notes.

  - Local Color versus Global Color. Node colors can be changed locally to a node or globally.
  - Label Inside Shape. Text can be placed inside the node (you may need to make the node wider to accommodate longer text).
  - Branch Event Name. Text can be placed on the branch leading to the node to indicate the event leading to this node.
  - Select Real Options. A specific real option type can be assigned to the current node. Assigning real options to nodes allows the tool to generate a list of required input variables.

- Global Elements are all customizable, including elements of the Strategy Tree's Background, Connection Lines, Option Nodes, Terminal Nodes, and Text Boxes. For instance, the following settings can be changed for each of the elements:
  - Font settings on Name, Value, Notes, Label, Event names.
  - Node Size (minimum and maximum height and width).
  - Borders (line styles, width, and color).
  - Shadow (colors and whether to apply a shadow or not).
  - Global Color.
  - Global Shape.
- The Edit menu's View Data Requirements Window command opens a docked window on the right of the Strategy Tree such that when an option node or terminal node is selected, the properties of that node will be displayed and can be updated directly. This feature provides an alternative to double-clicking on a node each time.
- Example Files are available in the File menu to help you get started on building Strategy Trees.
- Protect File from the File menu allows the Strategy Tree to be encrypted with up to a 256-bit password encryption. Be careful when a file is being encrypted because if the password is lost, the file can no longer be opened.
- Capturing the Screen or printing the existing model can be done through the File menu. The captured screen can then be pasted into other software applications.
- Add, Duplicate, Rename, and Delete a Strategy Tree can be performed through right-clicking the Strategy Tree tab or the Edit menu.
- You can also Insert File Link and Insert Comment on any option or terminal node, or Insert Text or Insert Picture anywhere in the background or canvas area.
- You can Change Existing Styles, or Manage and Create Custom Styles of your Strategy Tree (this includes size, shape, color schemes, and font size/color specifications of the entire Strategy Tree).
- Insert Decision, Insert Uncertainty, or Insert Terminal nodes by selecting any existing node and then clicking on the decision node icon (square), uncertainty node icon (circle), or terminal node icon (triangle), or use the functionalities in the Insert menu

- Modify individual Decision, Uncertainty, or Terminal nodes' properties by double-clicking on a node. The following are some additional unique items in the Decision Tree module that can be customized and configured in the node properties user interface:
    - Decision Nodes: Custom Override or Auto Compute the value on a node. The automatically compute option is set as default and when you click RUN on a completed Decision Tree model, the decision nodes will be updated with the results.
    - Uncertainty Nodes: Event Names, Probabilities, and Set Simulation Assumptions. You can add probability event names, probabilities, and simulation assumptions only after the uncertainty branches are created.
    - Terminal Nodes: Manual Input, Excel Link, and Set Simulation Assumptions. The terminal event payoffs can be entered manually or linked to an Excel cell (e.g., if you have a large Excel model that computes the payoff, you can link the model to this Excel model's output cell) or set probability distributional assumptions for running simulations.
- View Node Properties Window is available from the Edit menu and the selected node's properties will update when a node is selected.
- The Decision Tree module also comes with the following advanced analytics:
    - Monte Carlo Simulation Modeling on Decision Trees
    - Bayes Analysis for obtaining posterior probabilities
    - Expected Value of Perfect Information, MINIMAX and MAXIMIN Analysis, Risk Profiles, and Value of Imperfect Information
    - Sensitivity Analysis
    - Scenario Analysis
    - Utility Function Analysis

5.27.2 Simulation Modeling

This tool runs Monte Carlo risk simulation on the decision tree (Figure 5.62). It allows you to set probability distributions as input assumptions for running simulations. You can either set an assumption for the selected node or set a new assumption and use this new assumption (or use previously created assumptions) in a numerical equation or formula. For example, you can set a new assumption called Normal (e.g., normal distribution with a mean of 100 and standard deviation of 10) and run a simulation in the decision tree, or use this assumption in an equation such as (100*Normal+15.25).

Create your own model in the numerical expression box. You can use basic computations or add existing variables into your equation by double-clicking on the list of existing variables. New variables can be added to the list as required either as numerical expressions or assumptions.

5.27.3 Bayes Analysis

This Bayesian analysis tool (Figure 5.63) can be used on any two uncertainty events that are linked along a path. For instance, in the example on the right (Figure 5.63), uncertainties A and B are linked, where event A occurs first in the timeline and event B occurs second. First Event A is Market Research with 2 outcomes (Favorable or Unfavorable). Second Event B is Market Conditions also with 2 outcomes (Strong and Weak). This tool is used to compute joint, marginal, and Bayesian posterior updated probabilities by entering the prior probabilities and reliability conditional probabilities; or reliability probabilities can be computed when you have posterior updated conditional probabilities. Select the relevant analysis desired below and click on Load Example to see the sample inputs corresponding to the selected analysis and the results shown in the grid on the right, as well as which results are used as inputs in the decision tree in the figure.

Procedure

- STEP 1: Enter the names for the first and second uncertainty events and choose how many probability events (states of nature or outcomes) each event has.
- STEP 2: Enter the names of each probability event or outcome.
- STEP 3: Enter the second event's prior probabilities and the conditional probabilities for each event or outcome. The probabilities must sum to 100%.

5.27.4 Expected Value of Perfect Information, MINIMAX and MAXIMIN Analysis, Risk Profiles, and Value of Imperfect Information

This tool (Figure 5.64) computes the Expected Value of Perfect Information (EVPI), MINIMAX and MAXIMIN Analysis, as well as the Risk Profile and the Value of Imperfect Information. To get started, enter the number of decision branches or strategies under consideration (e.g., build a large, medium, or small facility), the number of uncertain events or states of nature outcomes (e.g., good market, bad market), and the expected payoffs under each scenario.

The Expected Value of Perfect Information (EVPI), that is, assuming you had perfect foresight and knew exactly what to do (through market research or other means to better discern the probabilistic outcomes), computes if there is added value in such information (i.e., if market research will add value) as compared to more naive estimates of the probabilistic states of nature. To get started, enter the number of decision branches or strategies under consideration (e.g., build a large, medium, or small facility) and the number of uncertain events or states of nature outcomes (e.g., good market, bad market), and enter the expected payoffs under each scenario.

MINIMAX (minimizing the maximum regret) and MAXIMIN (maximizing the minimum payoff) are two alternate approaches to finding the optimal decision path. These two approaches are not used often but still provide added insight into the decision-making process. Enter the number of decision branches or paths that exist (e.g., building a large, medium, or small facility), as well as the uncertainty events or states of nature under each path (e.g., good economy vs. bad economy). Then, complete the payoff table for the various scenarios and Compute the MINIMAX and MAXIMIN results. You can also click on Load Example to see a sample calculation.

5.27.5 Sensitivity

Sensitivity analysis (Figure 5.65) on the input probabilities is performed to determine the impact of inputs on the values of decision paths. First, select one Decision Node to analyze below, and then select one probability event to test from the list. If there are multiple uncertainty events with identical probabilities, they can be analyzed either independently or concurrently.

The sensitivity charts show the values of the decision paths under varying probability levels. The numerical values are shown in the results table. The location of crossover lines, if any, represents at what probabilistic events a certain decision path becomes dominant over another.

5.27.6 Scenario Tables

Scenario tables (Figure 5.66) can be generated to determine the output values given some changes to the input. You can choose one or more Decision paths to analyze (the results of each path chosen will be represented as a separate table and chart) and one or two Uncertainty or Terminal nodes as input variables to the scenario table.

Procedure

- Select one or more Decision paths to analyze from the list below.
- Select one or two Uncertainty Events or Terminal Payoffs to model.
- Decide if you wish to change the event's probability on its own or all identical probability events at once.
- Enter the input scenario range.

5.27.7 Utility Function Generation

Utility functions (Figure 5.67), or U(x), are sometimes used in place of expected values of terminal payoffs in a decision tree. U(x) can be developed two ways: using tedious and detailed experimentation of every possible outcome or an exponential extrapolation method (used here). They can be modeled for a decision maker who is risk-averse (downsides are more disastrous or painful than an equal upside potential), risk-neutral (upsides and downsides have equal attractiveness), or risk-loving (upside potential is more attractive). Enter the minimum and maximum expected value of your terminal payoffs and the number of data points in between to compute the utility curve and table.

If you had a 50:50 gamble where you either earn $X or lose -$X/2 versus not playing and getting a $0 payoff, what would this $X be? For example, if you are indifferent between a bet where you can win $100 or lose -$50 with equal probability compared to not playing at all, then your X is $100. Enter the X in the Positive Earnings box below. Note that the larger X is, the less risk-averse you are, whereas a smaller X indicates that you are more risk-averse.

Enter the required inputs, select the U(x) type, and click Compute Utility to obtain the results. You can also apply the computed U(x) values to the decision tree to re-run it, or revert the tree back to using expected values of the payoffs.

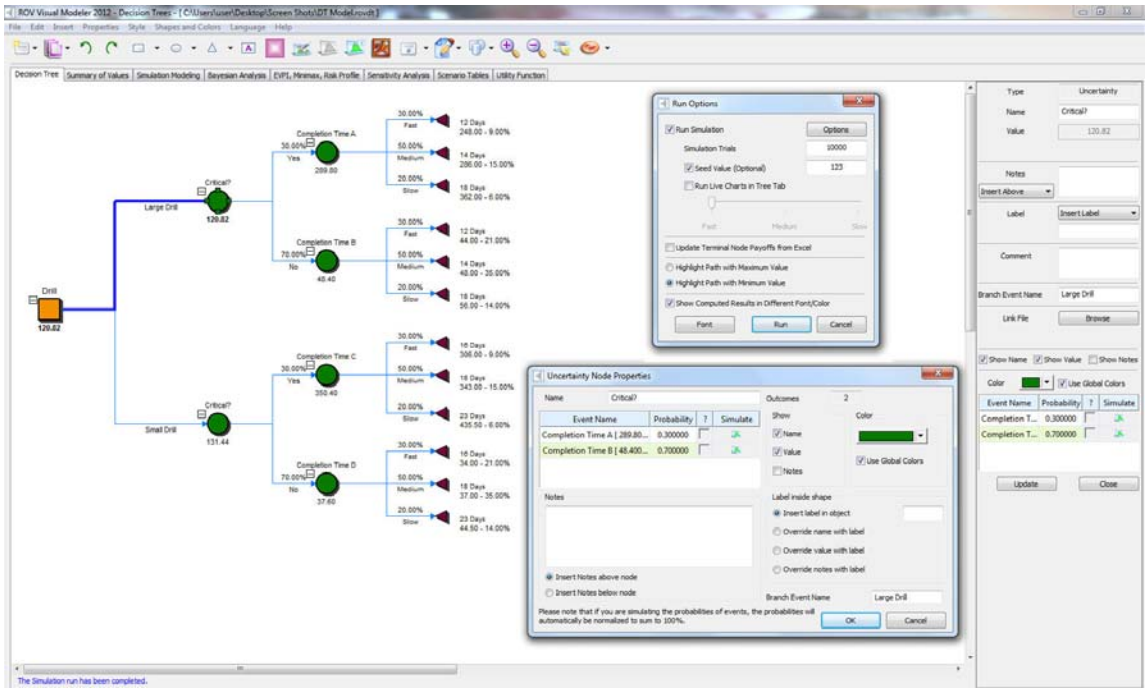


Figure 5.61 – ROV Decision Tree (Decision Tree)

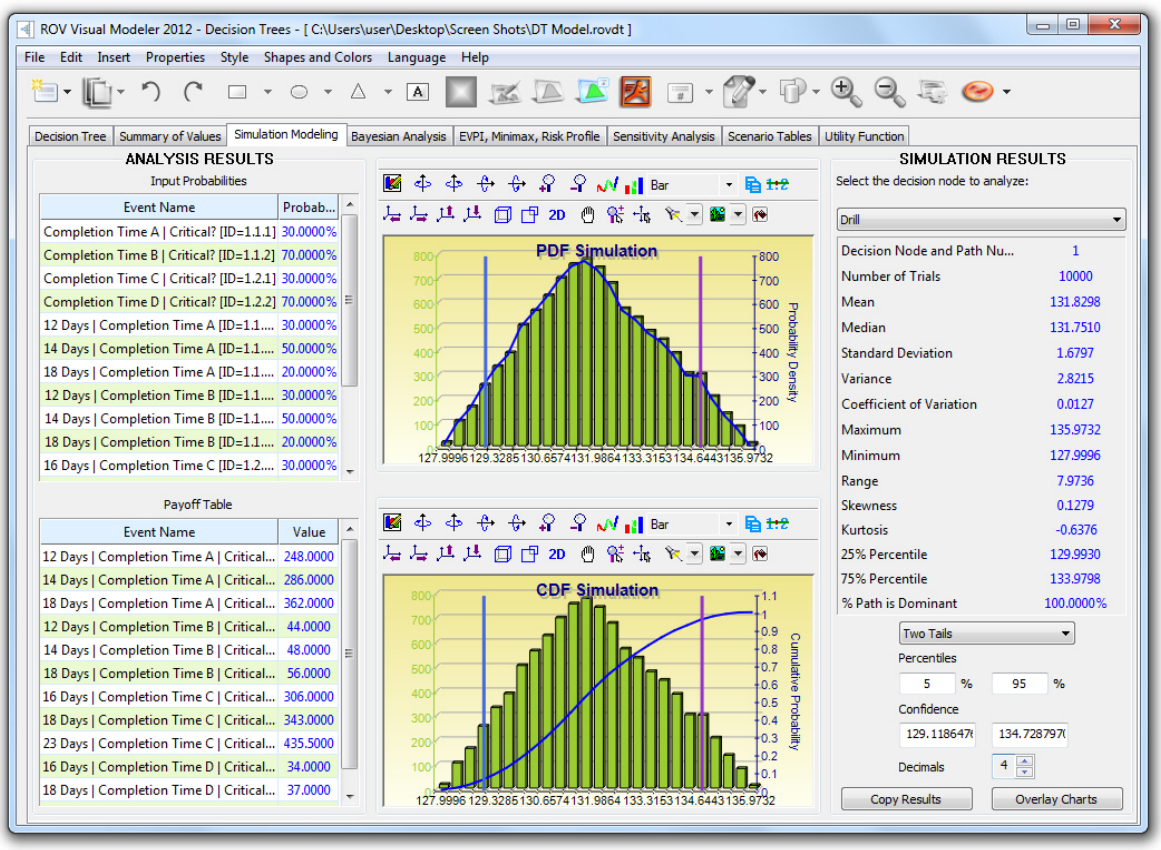


Figure 5.62 – ROV Decision Tree (Simulation Results)

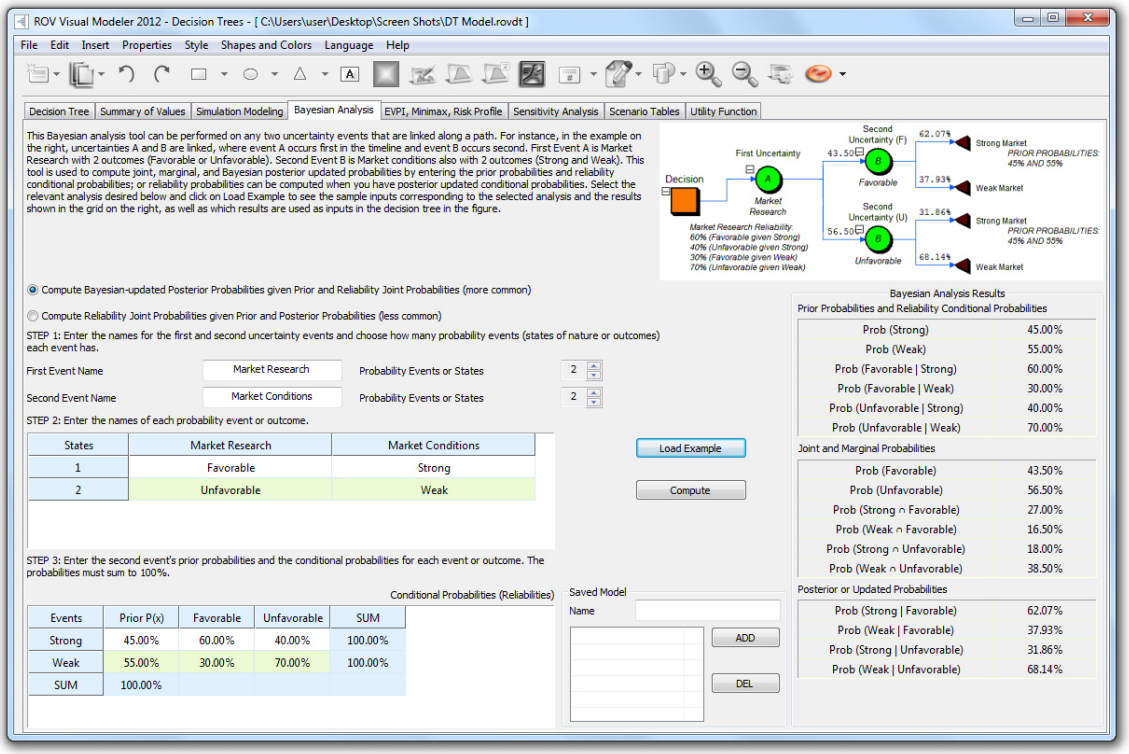


Figure 5.63 – ROV Decision Tree (Bayes Analysis)

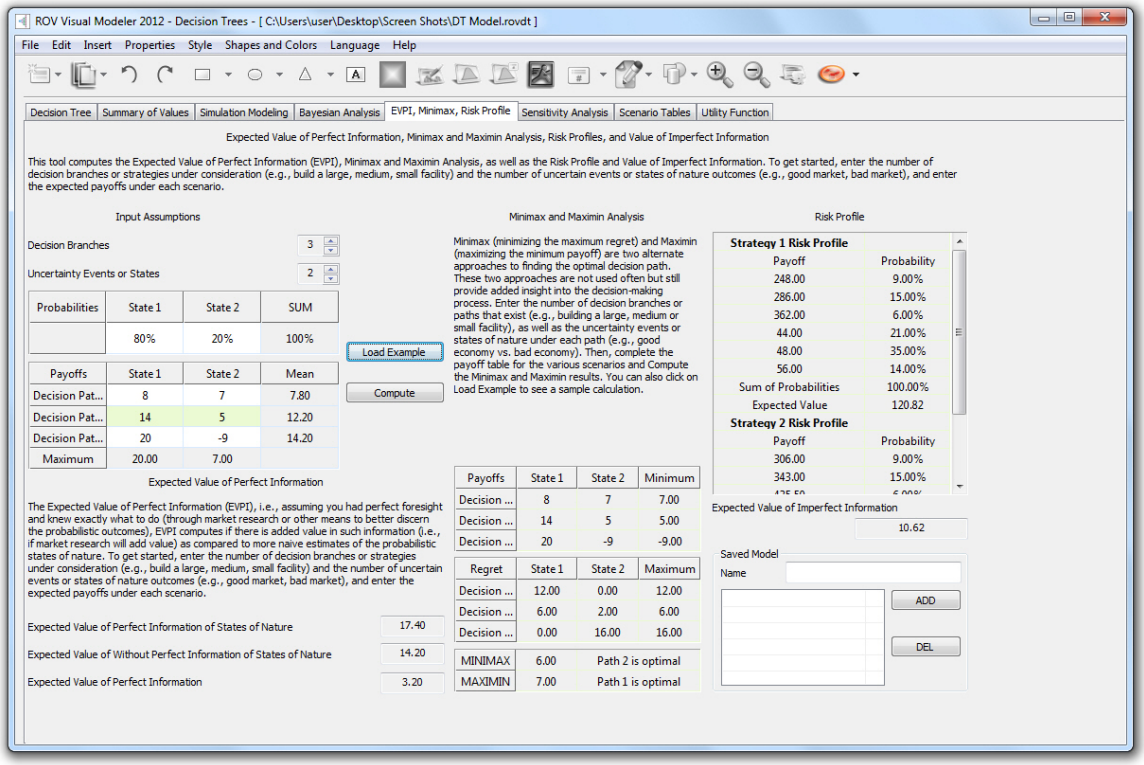


Figure 5.64 – ROV Decision Tree (EVPI, MINIMAX, Risk Profile)

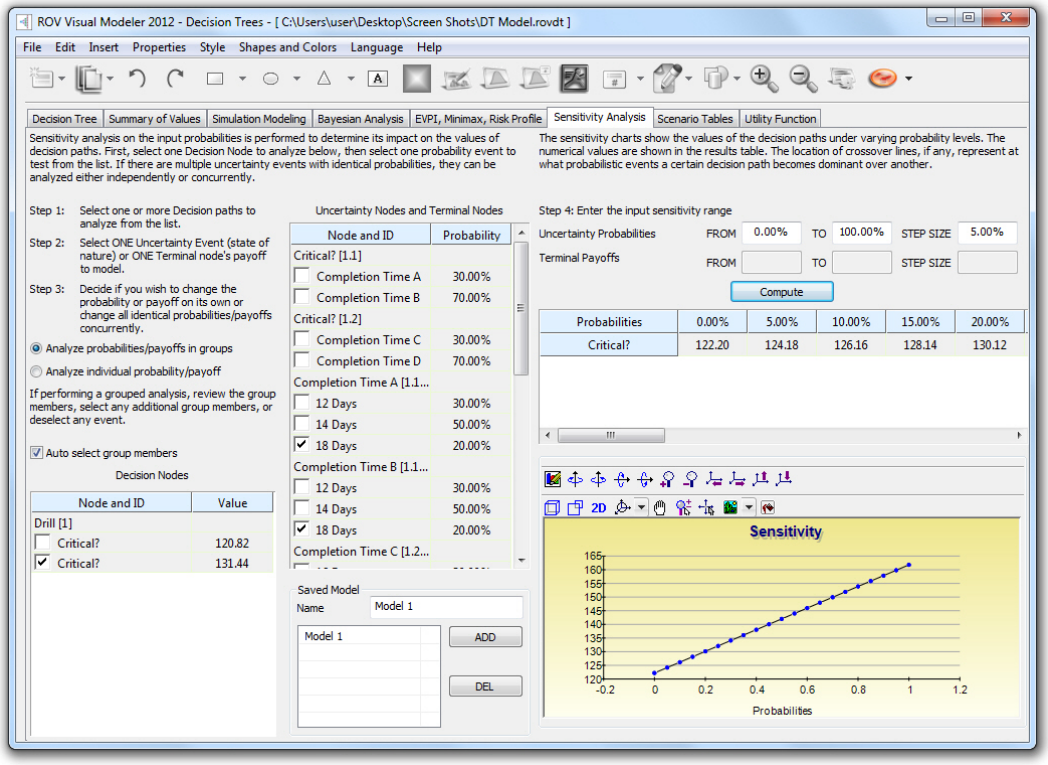


Figure 5.65 – ROV Decision Tree (Sensitivity Analysis)

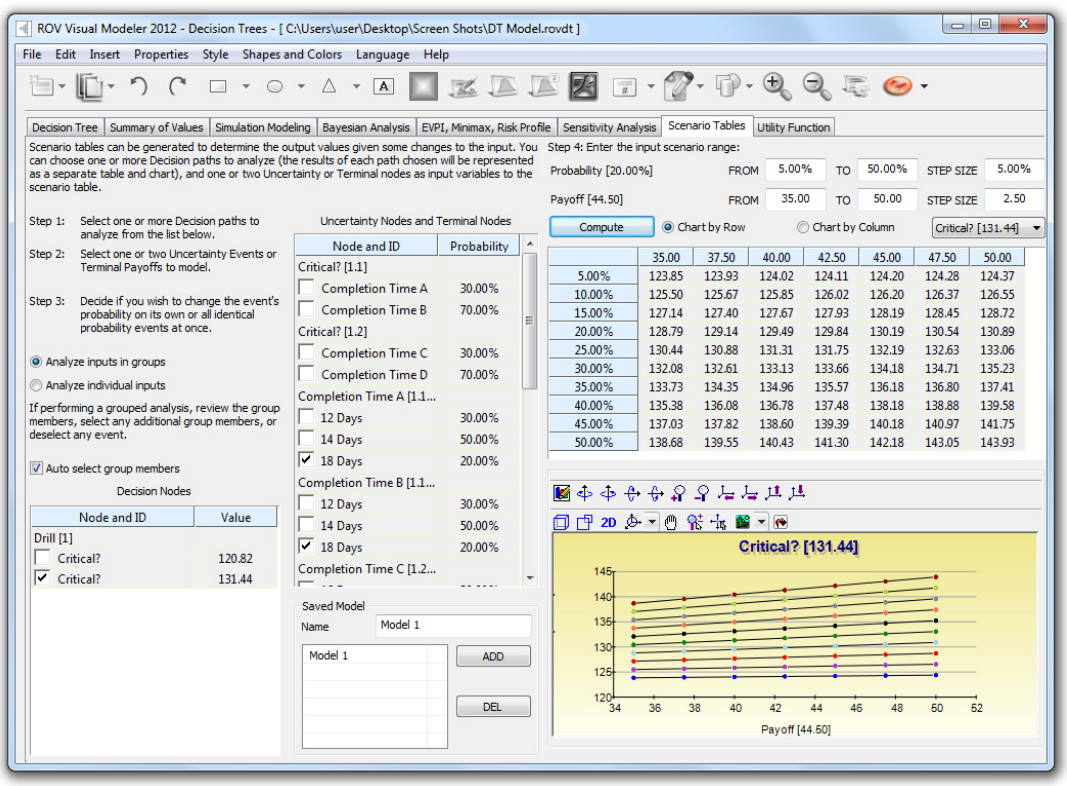


Figure 5.66 – ROV Decision Tree (Scenario Tables)

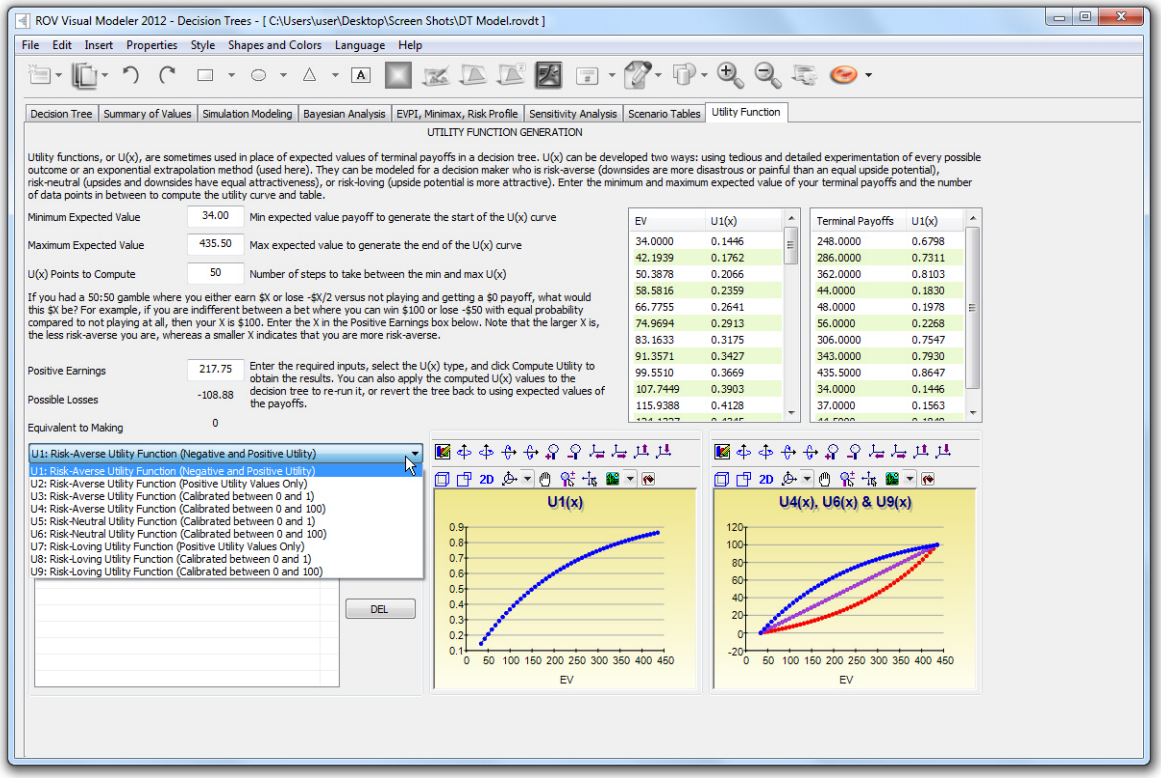


Figure 5.67 – ROV Decision Tree (Utility Functions)

6

6. Helpful Tips and Techniques

The following are some quick helpful tips and shortcut techniques for advanced users of Risk Simulator. For details on using specific tools, refer to the relevant sections in this user manual.

TIPS: Assumptions (Set Input Assumption User Interface)

- Quick Jump—select any distribution and type in any letter and it will jump to the first distribution starting with that letter (e.g., click on Normal and type in W and it will take you to the Weibull distribution).
- Right-Click Views—select any distribution, right-click, and select the different views of the distributions (large icons, small icons, list).
- Tab to Update Charts—after entering some new input parameters (e.g., you type in a new mean or standard deviation value), hit TAB on the keyboard or click anywhere on the user interface away from the input box to see the distributional chart automatically update.
- Enter Correlations—enter pairwise correlations directly here (the columns are resizable as needed), use the multiple distributional fitting tool to automatically compute and enter all pairwise correlations, or, after setting some assumptions, use the edit correlation tool to enter your correlation matrix.
- Equations in an Assumption Cell—only empty cells or cells with static values can be set as assumptions; however, there might be times when a function or equation is required in an assumption cell, and this can be done by first entering the input assumption in the cell and then typing in the equation or function (when the simulation is being run, the simulated values will replace the function, and after the simulation completes, the function or equation is again shown).

TIPS: Copy and Paste

- Copy and Paste using Escape—when you select a cell and use the Risk Simulator Copy function, it copies everything into Windows clipboard, including the cell's value, equation, function, color, font, and size, as well as Risk Simulator assumptions, forecasts, or decision variables. Then, as you apply the Risk Simulator Paste function, you have two options. The first option is to apply the Risk Simulator Paste directly, and all cell values, color, font, equation, functions and parameters will be pasted into the new cell. The second option is to first click

Escape on the keyboard, and then apply the Risk Simulator Paste. Escape tells Risk Simulator that you wish to paste only the Risk Simulator assumption, forecast, or decision variable, and not the cell's values, color, equation, function, font, and so forth. Hitting Escape before pasting allows you to maintain the target cell's values and computations, and pastes only the Risk Simulator parameters.

- Copy and Paste on Multiple Cells—select multiple cells for copy and paste (with contiguous and noncontiguous assumptions).

TIPS: Correlations

- Set Assumption—set pairwise correlations using the set input assumption dialog (ideal for entering only several correlations).
- Edit Correlations—set up a correlation matrix by manually entering or pasting from Windows clipboard (ideal for large correlation matrices and multiple correlations).
- Multiple Distributional Fitting—automatically computes and enters pairwise correlations (ideal when performing multiple variable fitting to automatically compute the correlations for deciding what constitutes a statistically significant correlation).

TIPS: Data Diagnostics and Statistical Analysis

- Stochastic Parameter Estimation—in the Statistical Analysis and Data Diagnostic reports, there is a tab on stochastic parameter estimations that estimates the volatility, drift, mean-reversion rate, and jump-diffusion rates based on historical data. Be aware that these parameter results are based solely on historical data used, and the parameters may change over time and depending on the amount of fitted historical data. Further, the analysis results show all parameters and do not imply which stochastic process model (e.g., Brownian Motion, Mean-Reversion, Jump-Diffusion, or mixed process) is the best fit. It is up to the user to make this determination depending on the time-series variable to be forecasted. The analysis cannot determine which process if best; only the user can do this (e.g., Brownian Motion process is best for modeling stock prices, but the analysis cannot determine that the historical data analyzed is from a stock or some other variable, and only the user will know this). Finally, a good hint is that if a certain parameter is out of the normal range, the process requiring this input parameter is most probably not the correct process (e.g., if the mean-reversion rate is 110%, chances are, mean-reversion is not the correct process).

TIPS: Distributional Analysis, Charts and Probability Tables

- Distributional Analysis—used to quickly compute the PDF, CDF, and ICDF of the 42 probability distributions available in Risk Simulator, and to return a table of these values.
- Distributional Charts and Tables—used to compare different parameters of the same distribution (e.g., takes the shapes and PDF, CDF, ICDF values of a Weibull

distribution with Alpha and Beta of [2, 2], [3, 5], and [3.5, 8] and overlays them on top of one another).

- Overlay Charts—used to compare different distributions (theoretical input assumptions and empirically simulated output forecasts) and overlay them on top of one another for a visual comparison.

TIPS: Efficient Frontier

- Efficient Frontier Variables—to access the frontier variables, first set the model's Constraints before setting efficient frontier variables.

TIPS: Forecast Cells

- Forecast Cells with No Equations—you can set output forecasts on cells without any equations or values (simply ignore the warning message) but be aware that the resulting forecast chart will be empty. Output forecasts are typically set on empty cells when there are macros that are being computed and the cell will be continually updated.

TIPS: Forecast Charts

- TAB versus Spacebar—hit TAB on the keyboard to update the forecast chart and to obtain the percentile and confidence values after you enter some inputs, and hit the Spacebar to rotate among the various tabs in the forecast chart.
- Normal versus Global View—click on these views to rotate between a tabbed interface and a global interface where all elements of the forecast charts are visible at once.
- Copy—copies the forecast chart or the entire global view depending on whether you are in the normal or global view.

TIPS: Forecasting

- Cell Link Address—if you first select the data in the spreadsheet and then run a forecasting tool, the cell address of the selected data will be automatically entered into the user interface Otherwise, you will have to manually enter in the cell address or use the link icon to link to the relevant data location.
- Forecast RMSE—use as the universal error measure on multiple forecast models for direct comparisons of the accuracy of each model.

TIPS: Forecasting: ARIMA

- Forecast Periods—the number of exogenous data rows has to exceed the time-series data rows by at least the desired forecast periods (e.g., if you wish to forecast 5 periods into the future and have 100 time-series data points, you will need to have

at least 105 or more data points on the exogenous variable). Otherwise, just run ARIMA without the exogenous variable to forecast as many periods as you wish without any limitations.

TIPS: Forecasting: Basic Econometrics

- Variable Separation with Semicolons—separate independent variables using a semicolon.

TIPS: Forecasting: Logit, Probit, and Tobit

- Data Requirements—the dependent variables for running logit and probit models must be binary only (0 and 1), whereas the Tobit model can take binary and other numerical decimal values. The independent variables for all three models can take any numerical value.

TIPS: Forecasting: Stochastic Processes

- Default Sample Inputs—when in doubt, use the default inputs as a starting point to develop your own model.
- Statistical Analysis Tool for Parameter Estimation—use this tool to calibrate the input parameters into the stochastic process models by estimating them from your raw data.
- Stochastic Process Model—sometimes if the stochastic process user interface hangs for a long time, chances are your inputs are incorrect and the model is not correctly specified (e.g., if the mean-reversion rate is 110%, mean-reversion is probably not the correct process). Try with different inputs or use a different model.

TIPS: Forecasting: Trendlines

- Forecast Results—scroll to the bottom of the report to see the forecasted values.

TIPS: Function Calls

- RS Functions—there are functions that you can use inside your Excel spreadsheet to set input assumption and get forecast statistics. To use these functions, you need to first install RS Functions (which include Start, Programs, Real Options Valuation, Risk Simulator, Tools, and Install Functions) and then run a simulation before setting the RS functions inside Excel. Refer to the example model 24 for examples on how to use these functions.

TIPS: Getting Started Exercises and Getting Started Videos

- Getting Started Exercises—there are multiple step-by-step hands-on examples and results interpretation exercises available in the Start, Programs, Real Options

Valuation, Risk Simulator shortcut location. These exercises are meant to quickly get you up to speed with the use of the software.

- Getting Started Videos—these are all available for free on our website: www.realoptionsvaluation.com/download.html or www.rovdownloads.com/download.html.

TIPS: Hardware ID

- Right-Click HWID Copy—in the Install License user interface, select or double-click on the HWID to select its value, right-click to copy or click on the E-mail HWID link to generate an e-mail with the HWID.
- Troubleshooter—run the Troubleshooter from the Start, Programs, Real Options Valuation, Risk Simulator folder, and run the Get HWID tool to obtain your computer's HWID.

TIPS: Latin Hypercube Sampling (LHS) vs. Monte Carlo Simulation (MCS)

- Correlations—when setting pairwise correlations among input assumptions, we recommend using the Monte Carlo setting in the Risk Simulator Options menu. Latin Hypercube Sampling is not compatible with the correlated copula method for simulation.
- LHS Bins—a larger number of bins will slow down the simulation while providing a more uniform set of simulation results.
- Randomness—all of the random simulation techniques in the Options menu have been tested and are all good simulators and approach the same levels of randomness when larger number of trials are run.

TIPS: Online Resources

- Books, Getting Started Videos, Models, White Papers—resources available on our website:

  www.realoptionsvaluation.com/download.html or www.rovdownloads.com/download.html.

TIPS: Optimization

- Infeasible Results—if the optimization run returns infeasible results, you can change the constraints from an Equal (=) to an Inequality (>= or <=) and try again. This also applies when you are running an efficient frontier analysis.

TIPS: Profiles

- Multiple Profiles—create and switch among multiple profiles in a single model. This allows you to run scenarios on simulation by being able to change input parameters or distribution types in your model to see the effects on the results.
- Profile Required—Assumptions, Forecasts, or Decision Variables cannot be created if there is no active profile. However, once you have a profile, you no longer have to keep creating new profiles each time. In fact, if you wish to run a simulation model by adding additional assumptions or forecasts, you should keep the same profile.
- Active Profile—the last profile used when you save Excel will be automatically opened the next time the Excel file is opened.
- Multiple Excel Files—when switching between several opened Excel models, the active profile will be from the current and active Excel model.
- Cross Workbook Profiles—be careful when you have multiple Excel files open because if only one of the Excel files has an active profile and you accidentally switch to another Excel file and set assumptions and forecasts on this file, the assumptions and forecast will not run and will be invalid.
- Deleting Profiles—you can clone existing profiles and delete existing profiles, but note that at least one profile must exist in the Excel file if you delete profiles.
- Profile Location—the profiles you create (containing the assumptions, forecasts, decision variables, objectives, constraints, etc.) are saved as an encrypted hidden worksheet. This is why the profile is automatically saved when you save the Excel workbook file.

TIPS: Right-Click Shortcut and Other Shortcut Keys

- Right-Click—you can open the Risk Simulator shortcut menu by right-clicking on a cell anywhere in Excel.

TIPS: Save

- Saving the Excel File—saves the profile settings, assumptions, forecasts, decision variables, and your Excel model (including any Risk Simulator reports, charts, and data extracted).
- Saving the Chart Settings—saves the forecast chart settings such that the same settings can be recovered and applied to future forecast charts (use the save and open icons in the forecast charts).
- Saving and Extracting Simulated Data in Excel—extracts a simulated run's assumptions and forecasts; the Excel file itself will still have to be saved in order to save the data for retrieval later.
- Saving Simulated Data and Charts in Risk Simulator—using the Risk Simulator Data Extract and saving to a *.RiskSim file will allow you to reopen the dynamic and

live forecast chart with the same data in the future without having to rerun the simulation.

- Saving and Generating Reports—simulation reports and other analytical reports are extracted as separate worksheets in your workbook, and the entire Excel file will have to be saved in order to save the data for future retrieval later.

TIPS: Sampling and Simulation Techniques

- Random Number Generator—there are six supported random number generators (see the user manual for details) and, in general, the ROV Risk Simulator default method and the Advanced Subtractive Random Shuffle method are the two recommended approaches to use. Do not apply the other methods unless your model or analytics specifically calls for their uses, and, even then, we recommended testing the results against these two recommended approaches.

TIPS: Software Development Kit (SDK) and DLL Libraries

- SDK, DLL, and OEM—all of the analytics in Risk Simulator can be called outside of this software and integrated in any user proprietary software. Contact admin@realoptionsvaluation.com for details on using our Software Development Kit to access the Dynamic Link Library (DLL) analytics files.

TIPS: Starting Risk Simulator with Excel

- ROV Troubleshooter—run this troubleshooter to obtain your computer's HWID for licensing purposes, to view your computer settings and prerequisites, and to re-enable Risk Simulator if it has been accidentally disabled.
- Start Risk Simulator when Excel Starts—you can let Risk Simulator start automatically when Excel starts each time or start it manually from the Start, Programs, Real Options Valuation, Risk Simulator shortcut location. This preference can be set in the Risk Simulator, Options menu.

TIPS: Super Speed Simulation

- Model Development—if you wish to run super speed in your model, test run a few super speed simulations while the model is being constructed to make sure that the final product will run the super speed simulation. Do not wait until the final model is complete before testing super speed to avoid having to backtrack to identify where any broken links or incompatible functions exist.
- Regular Speed—when in doubt, regular speed simulation always works.

TIPS: Tornado Analysis

- Tornado Analysis—the tornado analysis should never be run just once. It is meant as a model diagnostic tool, which means that it should ideally be run several times on

the same model. For instance, in a large model, Tornado can be run the first time using all of the default settings and all precedents should be shown (select Show All Variables). This single analysis may result in a large report and long (and potentially unsightly) Tornado charts. Nonetheless, it provides a great starting point to determine how many of the precedents are considered critical success factors. For example, the Tornado chart may show that the first 5 variables have high impact on the output, while the remaining 200 variables have little to no impact, in which case, a second tornado analysis is run showing fewer variables. For the second run, select Show Top 10 Variables if the first 5 are critical, thereby creating a nice report and a Tornado chart that shows a contrast between the key factors and less critical factors. (You should never show a Tornado chart with only the key variables without showing some less critical variables as a contrast to their effects on the output.)

- Default Values—the default testing points can be increased from the ±10% value to some larger value to test for nonlinearities (the Spider chart will show nonlinear lines and Tornado charts will be skewed to one side if the precedent effects are nonlinear).
- Zero Values and Integers—inputs with zero or integer values only should be deselected in the Tornado analysis before it is run. Otherwise, the percentage perturbation may invalidate your model (e.g., if your model uses a lookup table where Jan = 1, Feb = 2, Mar = 3, etc., perturbing the value 1 at a ±10% value yields 0.9 and 1.1, which makes no sense to the model).
- Chart Options—try various chart options to find the best options to turn on or off for your model.

TIPS: Troubleshooter

- ROV Troubleshooter—run this troubleshooter to obtain your computer's HWID for licensing purposes, to view your computer settings and prerequisites, and to re-enable Risk Simulator if it has been accidentally disabled.

INDEX

acquisition, 137, 156

allocation, 98, 99, 101, 102, 103, 104, 105, 110, 111

alpha, 45, 52, 69, 70, 126, 137

analysis, 1, 5, 9, 11, 12, 31, 33, 36, 51, 54, 67, 68, 69, 70, 72, 73, 74, 76, 78, 80, 81, 93, 98, 99, 110, 113, 115, 117, 119, 120, 121, 122, 124, 125, 133, 134, 137, 139, 140, 143, 149, 156, 168, 169, 172, 176, 177, 183, 186, 188, 189

annualized, 90, 100, 101, 110

approach, 11, 13, 26, 36, 64, 65, 66, 67, 72, 78, 80, 92, 93, 94, 98, 99, 100, 104, 110, 113, 119, 126, 138, 141, 153, 186

ARIMA, 1, 7, 9, 12, 65, 66, 68, 80, 81, 82, 84, 85, 86, 88, 89, 92, 95, 97, 138, 184

asset, 31, 99, 100, 101, 103, 104, 110, 111

asset classes, 100, 101, 110

assumption, 11, 14, 15, 16, 17, 19, 36, 37, 80, 92, 98, 99, 111, 113, 122, 124, 126, 137, 139, 175, 182, 183, 185

assumptions, 5, 6, 8, 11, 12, 13, 14, 15, 16, 17, 26, 27, 28, 37, 65, 67, 69, 70, 76, 88, 89, 91, 92, 98, 101, 102, 105, 107, 110, 111, 113, 115, 122, 123, 125, 126, 134, 135, 137, 139, 158, 161, 171, 175, 182, 183, 184, 186, 187

autocorrelation, 8, 66, 73, 81, 138, 140, 141

behavior, 45, 47, 66, 78, 86, 87, 126, 140

Beta, 45, 46, 47, 49, 51, 53, 54, 56, 57, 58, 62, 63, 161, 183

binomial, 6, 38, 39, 40, 41, 42, 43, 93, 146, 147, 161

Binomial, 38, 41, 147, 148, 149

bootstrap, 1, 12, 130, 131

Bootstrap, 8, 130, 131

Box-Jenkins, 1, 12, 66, 80, 82, 84, 85

Brownian Motion, 10, 140, 183

causality, 142

center of, 31, 118, 137
coefficient of determination, 136
confidence interval, 23, 24, 29, 60, 130, 132, 148
constraints, 12, 98, 99, 100, 102, 105, 108, 111, 171, 186, 187
Continuous, 36, 45, 98, 100, 103, 104, 111
correlation, 7, 8, 11, 14, 26, 27, 28, 37, 80, 81, 124, 138, 141, 142, 160, 182, 183
correlation coefficient, 14, 26, 27, 142
correlations, 8, 12, 14, 17, 26, 27, 28, 101, 122, 124, 142, 153, 182, 183, 186
cross-sectional, 65, 78
data, 6, 7, 8, 9, 12, 19, 21, 26, 27, 28, 35, 36, 45, 55, 62, 64, 65, 66, 67, 68, 69, 70, 72, 73, 76, 78, 80, 81, 85, 86, 87, 89, 90, 93, 95, 96, 97, 103, 126, 129, 130, 131, 133, 134, 136, 137, 138, 139, 140, 142, 143, 152, 153, 154, 155, 156, 157, 158, 159, 160, 164, 165, 168, 169, 177, 183, 184, 185, 187
decision variable, 5, 8, 12, 98, 99, 101, 102, 105, 110, 111, 113, 171, 182, 187
decision variables, 5, 12, 98, 99, 101, 102, 105, 110, 111, 113, 171, 182, 187
decisions, 98, 105, 113
Delphi, 6, 65, 126
Delphi method, 6, 126
dependent variable, 12, 66, 67, 72, 73, 80, 81, 85, 93, 94, 95, 136, 137, 138, 139, 185
discrete, 1, 6, 36, 37, 38, 39, 98, 99, 105, 129, 140, 146
Discrete, 36, 38, 39, 105, 106, 107
distribution, 6, 9, 11, 12, 15, 16, 17, 20, 26, 28, 31, 32, 33, 35, 36, 37, 38, 39, 40, 41, 42, 43, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 65, 66, 76, 91, 92, 94, 99, 113, 124, 126, 129, 130, 131, 139, 146, 147, 148, 153, 159, 160,161, 175, 182, 183, 186
Distribution, 6, 16, 27, 31, 32, 33, 36, 38, 40, 41, 42, 43, 45, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 91, 128, 129, 147, 148, 149, 150, 160, 161, 162, 163
distributional, 1, 8, 9, 11, 12, 13, 17, 21, 27, 31, 33, 36, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 66, 91, 111, 124, 125, 126, 129, 175, 182

distributions, 1, 6, 8, 9, 12, 16, 17, 26, 31, 33, 35, 36, 37, 38, 42, 46, 47, 52, 53, 56, 58, 62, 110, 126, 129, 130, 131, 132, 133, 146, 148, 153, 159, 160, 161, 175, 182, 183, 184

e-mail, 2, 3, 186

equation, 30, 34, 72, 76, 80, 87, 92, 116, 137, 140, 141, 175, 182

Erlang, 6, 49, 52

error, 1, 2, 14, 18, 21, 29, 56, 69, 72, 78, 80, 81, 94, 121, 130, 137, 138, 139, 155, 184

errors, 14, 19, 43, 66, 67, 72, 78, 80, 93, 136, 137, 139

estimates, 58, 65, 68, 78, 80, 93, 94, 126, 130, 137, 138, 176, 183

Excel, 1, 2, 3, 4, 5, 6, 8, 12, 13, 14, 16, 21, 26, 27, 34, 68, 69, 73, 78, 81, 85, 86, 88, 89, 92, 94, 95, 97, 98, 111, 116, 121, 165, 168, 173, 175, 185, 187, 188

excess kurtosis, 33, 38, 39, 40, 41, 42, 44, 46, 47, 50, 51, 52, 53, 54, 55, 56, 60, 61, 62

extrapolation, 1, 7, 9, 78, 177

first moment, 31, 34

Fisher-Snedecor, 51

flexibility, 113

fluctuations, 56, 122, 124, 136, 140, 154

forecast, 7, 8, 11, 14, 18, 19, 20, 21, 23, 25, 26, 28, 29, 30, 31, 36, 37, 65, 66, 67, 69, 70, 76, 78, 80, 81, 85, 86, 88, 89, 90, 92, 96, 98, 99, 111, 115, 123, 124, 130, 131, 132, 133, 137, 138, 140, 158, 168, 183, 184, 185, 187

forecast statistics, 8, 20, 98, 99, 130, 185

forecasting, 5, 7, 9, 11, 12, 36, 64, 65, 66, 67, 68, 70, 80, 85, 86, 88, 107, 136, 138, 140, 141, 154, 168, 169, 184

Forecasting, 1, 7, 9, 12, 64, 65, 68, 69, 70, 73, 74, 76, 77, 78, 81, 85, 86, 87, 88, 89, 92, 95, 97, 136, 138, 158, 168, 184, 185

forecasts, 5, 6, 8, 12, 13, 15, 17, 18, 26, 30, 65, 66, 68, 69, 70, 78, 80, 90, 98, 123, 132, 133, 134, 135, 138, 158, 161, 169, 182, 184, 187

fourth moment, 31, 33, 34

Frequency, 35, 36

functions, 1, 6, 7, 11, 15, 16, 18, 34, 36, 78, 87, 98, 105, 111, 121, 138, 153, 173, 177, 182, 185, 188

functions of, 138
gallery, 16, 17
gamma, 6, 47, 52, 60, 69, 70, 148
Gamma, 46, 49, 52, 57, 58, 62
geometric, 6, 40, 41, 55, 76, 100, 110
Geometric, 10, 38, 40
geometric average, 100, 110
goodness-of-fit, 138, 139
goodness-of-fit tests, 138
growth, 53, 65, 67, 80, 88, 110, 126, 140
growth rate, 65, 67, 80, 88, 126, 140
heteroskedasticity, 7, 8, 12, 66, 73, 89, 136, 137, 138, 139
Histogram, 21, 35, 36
Holt-Winter, 10, 69, 71
hypergeometric, 6, 40, 41
Hypergeometric, 40
hypothesis, 1, 12, 47, 51, 60, 74, 81, 94, 129, 130, 132, 133, 139, 142, 157
icon, 3, 6, 15, 16, 18, 19, 85, 89, 97, 101, 105, 111, 173, 174, 184
icons, 6, 16, 85, 111, 161, 165, 182, 187
independent variable, 66, 67, 72, 73, 81, 85, 86, 87, 93, 94, 137, 139, 141, 142, 156, 185
inflation, 55, 65, 138, 140, 142
inputs, 13, 15, 17, 27, 45, 46, 50, 53, 55, 61, 62, 66, 67, 68, 76, 85, 92, 98, 105, 111, 113, 115, 121, 149, 153, 160, 161, 172, 176, 177, 184, 185, 189
installation, 2, 3
integer, 1, 12, 14, 39, 41, 43, 47, 49, 52, 60, 67, 70, 87, 98, 99, 105, 121, 189
interest, 65, 66, 96, 138, 140, 149, 152
interest rate, 65, 66, 96, 138, 140
investment, 99, 105, 110, 117, 118, 119, 124
jump-diffusion, 2, 12, 67, 76, 140, 183
Kolmogorov-Smirnov test, 129
kurtosis, 33, 47, 56, 94

lags, 9, 73, 80, 81, 138
least squares, 73, 93, 94, 137
least squares regression, 93, 94, 137
linear, 7, 12, 26, 66, 68, 72, 80, 93, 96, 98, 100, 137, 139, 141, 142, 158
Ljung-Box Q-statistics, 81, 138
logistic, 6, 7, 12, 48, 53, 54, 67, 88, 93, 94
Lognormal, 54, 55
lower, 14, 27, 32, 40, 48, 54, 57, 60, 101, 110, 117, 124, 126
management, 58, 65, 113, 126, 152
market, 7, 26, 33, 65, 67, 88, 92, 126, 137, 140, 142, 152, 176
matrix, 26, 27, 141, 160, 182, 183
mean, 2, 11, 12, 17, 20, 21, 28, 29, 31, 32, 33, 36, 39, 46, 47, 53, 54, 55, 60, 66, 67, 76, 89, 92, 94, 99, 113, 126, 131, 137, 139, 140, 148, 155, 156, 175, 182, 183, 185
Mean, 9, 34, 39, 41, 47, 50, 51, 52, 53, 54, 55, 60, 61, 91, 140, 183
mean-reversion, 12, 67, 76, 140, 183, 185
mix, 141
model, 6, 7, 8, 11, 13, 14, 15, 17, 18, 19, 27, 28, 29, 36, 37, 57, 62, 65, 66, 67, 69, 70, 73, 76, 80, 81, 85, 86, 87, 88, 89, 92, 93, 94, 95, 97, 98, 99, 100, 101, 102, 105, 108, 110, 111, 113, 115, 116, 119, 120, 121, 122, 123, 124, 126, 134, 135, 136, 137, 138, 139, 140, 143, 149, 153, 158, 164, 165, 168, 170, 171, 172, 174, 175, 177, 183, 184, 185, 186, 187, 188, 189
Model, 8, 9, 13, 15, 18, 27, 28, 70, 89, 94, 95, 97, 100, 106, 111, 116, 149, 158, 159, 173, 185, 188
models, 2, 5, 6, 7, 9, 12, 14, 28, 58, 66, 67, 69, 70, 80, 81, 86, 87, 89, 90, 92, 94, 99, 105, 138, 154, 155, 164, 165, 169, 173, 184, 185, 187
Monte Carlo, 1, 5, 6, 11, 12, 13, 14, 17, 19, 26, 27, 29, 35, 36, 37, 65, 67, 70, 73, 74, 76, 81, 98, 153, 173, 175, 186
multicollinearity, 8, 12, 73, 136, 141, 142
Multinomial SLS, 2

multiple, 1, 2, 6, 8, 9, 11, 12, 13, 14, 18, 30, 37, 66, 67, 72, 73, 76, 81, 85, 87, 99, 110, 113, 121, 122, 126, 129, 141, 143, 149, 156, 160, 161, 164, 165, 177, 182, 183, 184, 185, 186, 187
multiple regression, 1, 67, 141
multiple variables, 8, 87, 129, 143, 156, 160, 164, 165
multivariate, 8, 12, 72, 73, 74, 78, 80, 81, 93, 153
Mun, 0, i, 1, 65, 70, 73, 74, 76, 81, 89, 90, 107
negative binomial, 6, 41, 42, 43, 148
nonlinear, 1, 7, 9, 12, 66, 67, 68, 73, 78, 80, 96, 98, 99, 115, 121, 124, 125, 137, 142, 158, 189
normal, 6, 11, 17, 26, 30, 33, 36, 39, 47, 53, 54, 55, 58, 60, 76, 91, 92, 93, 94, 126, 130, 137, 148, 153, 175, 183, 184
Normal, 6, 21, 55, 91, 94, 150, 153, 175, 182, 184
null hypothesis, 81, 126, 133, 137, 138, 139, 157
objective, 12, 98, 99, 100, 101, 102, 105, 111
optimal, 98, 99, 103, 104, 107, 108, 113, 137, 176
optimal decision, 99, 113, 176
optimization, 1, 7, 8, 9, 12, 66, 89, 93, 98, 99, 100, 101, 102, 103, 104, 105, 107, 108, 110, 111, 113, 129, 171, 172, 186
option, 2, 17, 69, 70, 120, 121, 126, 134, 161, 165, 172, 173, 174, 175, 182
outliers, 8, 136, 137, 138, 139
parameter, 16, 17, 38, 40, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 59, 60, 62, 63, 69, 92, 94, 121, 141, 164, 168, 183
Parameter, 55, 111, 141, 146, 183, 185
pareto, 56
Pareto, 6, 56, 57
pause, 18, 19
Pearson, 6, 26, 27, 57, 58, 142
point estimate, 65, 69, 70, 99, 113
Poisson, 6, 43, 44, 50, 52
population, 29, 34, 41, 53, 56, 60, 67, 88, 126, 132, 133, 137, 139, 156

portfolio, 1, 98, 99, 100, 101, 102, 103, 104, 105, 110, 113

precision, 1, 8, 14, 18, 21, 29

prediction, 80, 137, 138

price, 27, 31, 54, 55, 66, 76, 89, 117, 118, 119, 122, 140, 141, 149

probability, 1, 6, 8, 9, 11, 19, 23, 24, 32, 35, 36, 37, 38, 39, 40, 41, 42, 43, 45, 47, 48, 51, 53, 55, 58, 59, 60, 67, 92, 93, 94, 146, 147, 148, 159, 160, 161, 175, 176, 177, 183

Probability, 5, 9, 17, 25, 35, 36, 38, 39, 40, 41, 42, 43, 146, 161, 162, 163, 173, 183

profile, 13, 14, 15, 27, 69, 101, 105, 111, 126, 164, 165, 186, 187

p-value, 81, 126, 133, 138, 142, 157

random, 6, 11, 12, 14, 26, 35, 36, 37, 45, 46, 47, 50, 52, 53, 67, 76, 80, 94, 130, 131, 132, 140, 141, 153, 161, 186, 188

random number, 6, 11, 14, 36, 153, 188

range, 17, 31, 45, 52, 93, 99, 101, 110, 113, 115, 138, 147, 177, 183

rank correlation, 26, 27, 124, 125, 142

rate, 7, 31, 44, 49, 50, 55, 66, 76, 94, 96, 117, 119, 124, 138, 140, 149, 155, 183, 185

ratio, 47, 51, 78, 94, 99, 101, 103, 104, 105, 107, 110

regression, 7, 66, 67, 72, 73, 74, 78, 80, 81, 85, 86, 93, 94, 136, 137, 138, 139, 140, 141

Regression, 7, 9, 65, 67, 68, 72, 73, 74, 75, 80, 86, 136

regression analysis, 66, 67, 72, 81, 85, 86, 93, 94, 136, 137

relative returns, 66, 89, 100, 110

report, 8, 14, 69, 74, 76, 78, 81, 86, 87, 88, 89, 92, 95, 96, 97, 108, 111, 117, 121, 124, 126, 135, 136, 138, 140, 142, 155, 156, 157, 158, 164, 165, 185, 188

return, 31, 67, 92, 93, 100, 101, 103, 104, 110, 121, 161, 172, 183

returns, 31, 32, 33, 66, 76, 85, 99, 100, 101, 102, 103, 104, 105, 110, 111, 137, 153, 186

risk, 5, 8, 11, 12, 28, 31, 32, 33, 34, 36, 52, 67, 100, 101, 102, 103, 104, 105, 107, 110, 111, 115, 119, 172, 173, 175, 177

Risk Simulator, 1, 2, 3, 4, 5, 6, 12, 13, 14, 15, 16, 17, 18, 19, 25, 26, 27, 29, 30, 31, 33, 37, 65, 66, 67, 68, 69, 70, 73, 76, 78, 81, 85, 86, 87, 88, 89, 90, 91, 92, 94, 95, 97, 98, 99, 100, 101, 102, 103, 105, 107, 108, 111, 115, 116, 120, 123, 126, 129, 130,

132, 134, 135, 136, 138, 143, 146, 148, 149, 152, 153, 154, 155, 156, 157, 158, 160, 161, 164, 168, 169, 170, 171, 172, 173, 182, 183, 185, 186, 187, 188, 189
running, 5, 6, 14, 19, 21, 28, 73, 87, 98, 108, 119, 120, 130, 137, 139, 153, 164, 169, 175, 185, 186
sales, 41, 42, 61, 65, 67, 68, 72, 88, 138, 154
sample, 11, 29, 34, 35, 41, 60, 69, 74, 76, 87, 88, 100, 107, 110, 115, 126, 130, 131, 132, 137, 139, 143, 149, 152, 153, 161, 164, 176
save, 3, 14, 134, 164, 187
saving, 5, 134, 187
seasonality, 9, 12, 68, 69, 70, 138, 154, 155, 168, 169
second moment, 31, 33, 34, 159
sensitivity, 1, 9, 12, 117, 119, 122, 124, 125, 177
Sensitivity, 5, 9, 115, 119, 120, 122, 123, 124, 173, 175, 177, 180
significance, 8, 60, 81, 94, 130, 133, 137, 138, 139, 142, 157
simulation, 1, 5, 6, 7, 8, 11, 12, 13, 14, 15, 17, 18, 19, 21, 25, 26, 27, 28, 29, 35, 36, 37, 65, 67, 69, 76, 98, 99, 101, 102, 105, 107, 110, 111, 113, 115, 119, 122, 123, 124, 125, 126, 130, 131, 132, 134, 135, 153, 158, 171, 173, 175, 182, 185, 186, 187, 188
Simulation, 1, 6, 7, 8, 11, 12, 13, 14, 15, 17, 18, 19, 25, 27, 35, 36, 65, 68, 70, 73, 74, 76, 81, 98, 110, 113, 115, 130, 131, 134, 135, 153, 175, 178, 186, 188
single, 6, 8, 14, 19, 21, 65, 67, 69, 70, 72, 87, 99, 113, 116, 126, 137, 143, 153, 160, 164, 170, 171, 186, 188
Single Asset SLS, 2
skew, 31, 32, 34
Skew, 32, 33, 34
skewness, 32, 33, 38, 39, 40, 41, 42, 44, 46, 47, 50, 51, 52, 53, 54, 55, 56, 60, 61, 62, 130
SLS, 1, 2, 6
Spearman, 26, 27, 142
specification errors, 136
spider, 1, 9, 12, 116, 117, 120, 122
spread, 28, 31

standard deviation, 11, 17, 21, 28, 30, 31, 32, 33, 34, 36, 39, 41, 44, 46, 47, 50, 51, 52, 54, 55, 60, 61, 76, 94, 98, 99, 113, 126, 131, 133, 139, 142, 148, 175, 182
static, 9, 12, 98, 101, 102, 105, 108, 115, 119, 122, 140, 182
statistics, 1, 8, 9, 12, 19, 20, 21, 26, 28, 29, 31, 34, 81, 98, 99, 113, 126, 130, 131, 143, 164
stochastic, 1, 8, 9, 12, 67, 76, 98, 99, 102, 105, 107, 108, 110, 111, 112, 113, 114, 136, 140, 143, 172, 183, 185
stochastic optimization, 8, 12, 98, 99, 102, 105, 110, 111, 112, 113, 114, 172
stock price, 31, 56, 66, 76, 89, 140, 183
symmetric, 137
t distribution, 60, 91, 92
third moment, 31, 32, 34
time-series, 1, 7, 9, 12, 65, 66, 67, 68, 69, 70, 76, 78, 80, 81, 85, 87, 96, 138, 140, 154, 156, 158, 168, 169, 183, 184
time-series data, 9, 66, 67, 68, 78, 81, 85, 87, 96, 138, 140, 154, 156, 158, 168, 184
title, 13, 14
toolbar, 3, 6, 15, 18, 19
tornado, 1, 9, 12, 115, 116, 117, 118, 119, 120, 122, 124, 125, 188
Tornado, 9, 115, 116, 117, 118, 119, 120, 121, 122, 123, 125, 188, 189
trends, 7, 78, 140
trials, 7, 11, 14, 18, 19, 29, 37, 38, 39, 40, 41, 42, 43, 50, 98, 99, 111, 113, 130, 147, 161, 186
triangular, 6, 11, 36, 58, 61
Triangular, 15, 58, 61
t-statistic, 94, 141
types of, 26, 37, 100, 110, 129, 140, 158, 160
uniform, 6, 11, 33, 36, 39, 61, 62, 101, 110, 126, 186
Uniform, 17, 39, 61
upper, 101, 110
validity of, 78, 138

value, 5, 8, 14, 17, 19, 23, 24, 26, 28, 29, 31, 33, 34, 36, 37, 38, 39, 45, 46, 47, 48, 50, 51, 52, 53, 54, 55, 57, 58, 59, 60, 61, 62, 63, 66, 67, 68, 70, 78, 80, 88, 90, 94, 101, 105, 108, 110, 113, 115, 117, 118, 121, 124, 126, 130, 131, 137, 138, 139, 140,141, 142, 146, 147, 148, 149, 154, 165, 168, 170, 171, 173, 175, 176, 177, 182, 185, 186, 189

values, 11, 14, 15, 19, 20, 21, 23, 24, 25, 26, 27, 28, 33, 36, 37, 45, 46, 48, 51, 54, 56, 58, 61, 65, 66, 67, 68, 69, 76, 78, 80, 81, 88, 93, 94, 96, 97, 98, 99, 100, 101, 102, 105, 107, 110, 115, 117, 121, 126, 133, 134, 137, 138, 139, 142, 146, 147, 149, 153, 154, 157, 161, 168, 177, 182, 183, 184, 185, 189

variance, 31, 32, 47, 51, 54, 57, 60, 74, 92, 113, 133, 136, 137, 142

volatility, 7, 12, 66, 76, 89, 90, 140, 183

Weibull, 6, 62, 63, 124, 161, 182, 183

Yes/No, 38

© Copyright 2005-2012 Dr. Johnathan Mun
All rights reserved.
Real Options Valuation, Inc.
4101F Dublin Blvd., Ste. 425
Dublin, California 94568 U.S.A.
Phone 925.271.4438 • Fax 925.369.0450
admin@realoptionsvaluation.com
www.risksimulator.com
www.realoptionsvaluation.com